Maria Manuel Costa Magalhães Teixeira

Porto de Aveiro: Um Estudo sobre a Exportação de Cimento

Maria Manuel Costa Magalhães Teixeira

Porto de Aveiro: Um Estudo sobre a Exportação de Cimento

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática Aplicada à Estatística e Investigação Operacional, realizado sob a orientação científica da Doutora Isabel Maria Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Em memória da minha avó materna, que sempre me acompanhou e sempre acreditou que eu iria conseguir.

o júri

presidente

Prof. Doutor Agostinho Miguel Mendes Agra
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

vogais

Prof. Doutora Anabela Virgínia dos Santos Flores da Rocha
Professora Adjunta do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro

Prof. Doutora Isabel Maria Simões Pereira
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

agradecimentos

Seguem-se os meus mais humildes e sinceros agradecimentos.

Primeiro que tudo, à minha orientadora da Universidade de Aveiro, a Professora Doutora Isabel Maria Simões Pereira, pela sua orientação e por todos os conhecimentos que me transmitiu ao longo do meu percurso académico.

À Administração do Porto de Aveiro, S.A., em especial ao Dr. Luís Sousa, pela disponibilidade e oportunidades que me deram. A todos os colaboradores da empresa que me fizeram sentir mais integrada e que me apoiaram ao longo do tempo de estágio deixo um grande agradecimento.

À minha colega de estágio, Maryse Oliveira, pelo seu companheirismo e apoio.

Aos meus pais pelo seu amor e compreensão. Estou-lhes eternamente grata pela oportunidade que me deram em realizar o meu percurso académico e por todos os esforços que fizeram para que a minha experiência fosse notável.

À minha irmã pela maneira entusiasta como sempre me apoiou, por toda a disponibilidade e energia que me transmitiu sempre. Agradeço-te imenso.

Ao meu extraordinário namorado, Diogo Amaral, pelo seu amor e pela sua incansável paciência ao longo do estágio e do processo de escrita. Um especial agradecimento também à sua família por todo o apoio e carinho.

A todos os meus amigos que acreditaram em mim, me ajudaram e apoiaram em todos os momentos destes últimos cinco anos.

palavras-chave Administração Portuária, Porto de Aveiro, Exportação de Cimento, Modelos de Regressão Linear Múltipla, Árvores de Regressão.

resumo O presente relatório, realizado no âmbito do curso de Mestrado de Matemática Aplicada à Estatística e Investigação Operacional, da Universidade de Aveiro, pretende retratar o estágio curricular realizado na Administração do Porto de Aveiro, S.A., com foco na exportação de cimento do Porto de Aveiro.

Utilizando Modelos de Regressão Linear Múltipla e Árvores de Regressão efetuou-se um estudo estatístico sobre a exportação de cimento do Porto de Aveiro. Para os Modelos de Regressão Linear Múltipla foi utilizado o *software* SPSS e para as Árvores de Regressão o *software* R.

keywords Port Administration, Aveiro Port, Exportation of Cement, Multiple Linear Regression Models, Regression Trees.

abstract This report, conducted under the master degree on Applied Mathematics to Statistics and Operational Research, from Aveiro University, aims to represent the traineeship held in Administração do Porto de Aveiro, S.A., with a focus on the export of cement from the Port of Aveiro.

Using Multiple Linear Regression Models and Regression Trees a statistical study of the export of cement from the Port of Aveiro was performed. For the Multiple Linear Regression Model was used software SPSS and for the Regression Trees software R.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

APA Administração do Porto de Aveiro, S.A.

APFF Administração do Porto da Figueira da Foz, S.A.

JARBA Junta Autónoma da Ria e Barra da Aveiro

JAPA Junta Autónoma do Porto de Aveiro

RO-RO *Roll on Roll Off*

Z.H. Zero Hidrográfico

df Graus de Liberdade

Capítulo 1

Introdução

O presente relatório de estágio está inserido no plano curricular do segundo ano do mestrado de Matemática Aplicada à Estatística e Investigação Operacional, lecionado na Universidade de Aveiro (UA). O estágio teve lugar na APA - Administração do Porto de Aveiro, S.A. com uma duração de seis meses, tendo início a 20 de janeiro de 2014 e término a 18 de julho de 2014.

A escolha do Porto de Aveiro para a realização do estágio foi uma oportunidade de ganhar conhecimento sobre o setor portuário sobre o qual não tinha qualquer domínio. Os portos sempre possuíram um papel importante no desenvolvimento do comércio dos países, tanto a nível nacional como internacional. Hoje em dia fortificado pela globalização, são fundamentais para o crescimento sustentado da economia das regiões onde estão inseridos. Daqui surgiu curiosidade sobre o setor.

A função principal de um porto é fornecer benefícios para os proprietários das cargas ou mercadorias e para o cliente final, ou seja para a população. Assim criando mais negócios, mais emprego e mais crescimento económico, sendo um pilar de desenvolvimento nacional e regional.

Estes fenómenos de desenvolvimento e de crescimento têm vindo a obrigar os portos a investir cada vez mais nas suas infra-estruturas, no aprofundamento dos seus cais e dos seus canais de navegação, para poderem acolher novos navios.

O Porto de Aveiro tem vindo a crescer nos últimos anos e a atingir totais de mercadoria movimentada mais elevados ano após ano. Com o surgimento da ferrovia em 2010 notou-se um claro crescimento na exportação de cimento. O cimento que tem vindo a liderar a lista de mercadorias movimentadas transportadas através do Porto de Aveiro, segundo

a administração do porto. Com este intuito um dos objetivos deste relatório é estudar a exportação de cimento desde o seu início, comprovando que variáveis mais influenciam este aumento na exportação.

Este estudo foi efetuado em paralelo com as responsabilidades diárias exigidas pela APA, S.A. As atividades desenvolvidas durante o estágio centraram-se sempre na estatística portuária, por forma a desenvolver melhorias na implementação e análise dos dados recolhidos pelos sistemas de informação. A estatística portuária é uma estatística bastante descritiva, pelo que é crucial a criação de novos indicadores que beneficiem os tratamentos estatísticos.

O relatório encontra-se estruturado em 5 capítulos. O presente capítulo, introduz os objetivos do estágio e a organização interna do relatório. No capítulo 2 é caracterizada a organização de acolhimento, desde a sua história, perfil, localização, área de intervenção e organograma. Este capítulo conta também com uma descrição das atividades desenvolvidas durante o estágio.

No capítulo 3 é apresentada a metodologia do estudo, expondo os procedimentos utilizados para a sua concretização. No quarto capítulo é apresentado o problema, os dados e a análise dos resultados obtidos. Este capítulo é suportado teoricamente pelo terceiro capítulo.

No último capítulo, é feita uma reflexão crítica sobre os resultados obtidos e do seu interesse para a APA, S.A., são apresentadas as considerações finais sobre o estudo também como as suas limitações.

Capítulo 2

Enquadramento da Empresa

Este capítulo tem como objetivo apresentar a organização de acolhimento onde o estágio se realizou. Para isso, é feita uma breve descrição da evolução histórica do porto de Aveiro e são apresentadas as zonas portuárias, as acessibilidades e o organograma. Neste capítulo encontra-se também uma apresentação resumida do porto da Figueira da Foz e são descritas as atividades desenvolvidas ao longo do estágio.

Para a realização desta secção, foram analisadas as referências [22], [19], [23], [20], [21] e [1].

2.1 Porto de Aveiro

2.1.1 Organização de Acolhimento

O Porto de Aveiro é administrado pela APA - Administração do Porto de Aveiro, S.A. (APA, S.A.), sociedade anónima de Capital Social exclusivamente público, detido pelo estado através da Direção Geral do Tesouro, que visa a exploração económica, conservação e desenvolvimento do Porto de Aveiro.

O Edifício da sede da APA (ver figura 2.1), localiza-se no Forte da Barra, numa zona administrativa onde coexiste com uma série de outros edifícios ocupados por autoridades e empresas que trabalham diretamente com o Porto de Aveiro.

Figura 2.1: Sede da APA, S.A.

2.1.2 Evolução Histórica

A origem do Porto de Aveiro enleia-se à história da ria de Aveiro e à obra de fixação e abertura da Barra. Com a finalidade de criar uma ligação do mar à ria de Aveiro, a cidade testemunhou constantes intervenções político-económicas e também técnicas. Segundo a história, desde 1757 foi realizada uma panóplia de estudos técnicos para a fixação do acesso marítimo.

O primeiro grande estudo para a criação da Barra do Porto de Aveiro, datado em 3 de abril de 1808, foi concebido pelos Engenheiros Reinaldo Oudinot e Luís Gomes de Carvalho. É de notar que a abertura da barra firmou-se como o principal símbolo do desenvolvimento do Porto de Aveiro. A identidade de Aveiro acabou então por se fundir com a abertura do porto ao comércio internacional. O setor da pesca, tanto costeira como longínqua, atraiu indústrias, tornando-se um dos mais importantes polos desta atividade a nível nacional e reabilitando o crescimento económico da região.

Até meados do século XX, deu-se continuidade às obras, ampliando-se molhes e construindo-se diques. A projeção de um porto de pesca e de um porto comercial junto ao canal de S. Roque constitui um dos primeiros planos de investimento para o Porto de Aveiro, cujo autor foi o engenheiro Von Hafe. Em meados do século XX, é criada a Junta Autónoma da Ria e Barra de Aveiro (JARBA). Foram também criados o ”Esquema Geral do Porto Interior de Aveiro”e os planos de arranjo e exploração do porto bacalhoeiro, do porto de pesca costeira e do porto comercial, por orientação do Engenheiro Coutinho de Lima.

O "Plano Diretor de Desenvolvimento e Valorização do Porto e Ria de Aveiro", em 1974, gerou uma deslocação dos terminais portuários para uma região próxima da entrada da Barra. Aquando deste plano, a JARBA já havia se transformado em JAPA (Junta Autónoma do Porto de Aveiro).

Em 1998, deu-se a conversão da JAPA em APA - Administração do Porto de Aveiro, S.A. Desde então, é reconhecido ao porto o estatuto de um porto nacional. A agora denominada APA originou uma reforma do "Plano de Ordenamento e Expansão do Porto de Aveiro", englobando a ligação ferroviária do Porto de Aveiro à linha do norte e a conclusão e melhoria das infraestruturas. Tal revisão apenas foi possível devido ao desenvolvimento do porto que conferiu à APA novas competências.

Em 2005, a área de jurisdição da APA ficou restrita aos espaços com interesse portuário. Consequentemente, de forma a garantir a gestão e futuro do porto até 2015, foi elaborado o "Plano Estratégico do Porto de Aveiro".

Na atualidade, a APA, S.A. tem como missão "facultar o acesso competitivo de mercadorias aos mercados regionais, nacionais e internacionais, promovendo assim o desenvolvimento económico da sua região". (ver [22] e [19])

2.1.3 Zonas Portuárias

O Porto de Aveiro é um porto multifuncional e apresenta uma área bem ordenada e integrada, sem congestionamentos, dispondo de 7 terminais especializados e 2 zonas logísticas intermodais (ver figura 2.2).

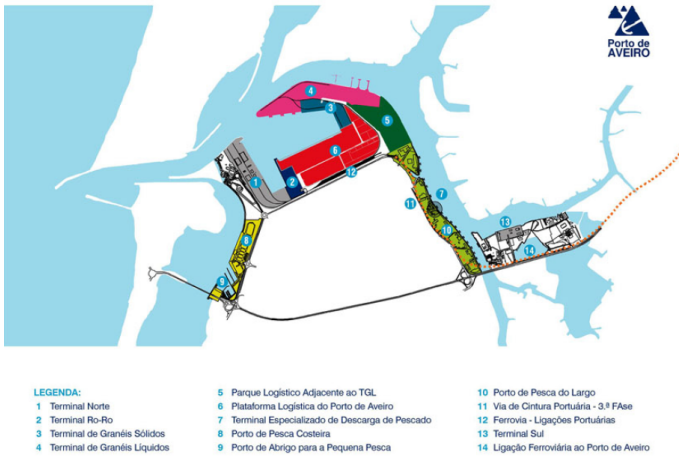

Figura 2.2: Planta do Porto de Aveiro.

Seguidamente, identificarei os terminais do Porto de Aveiro e também as suas zonas logísticas, especificando cada um deles.

Terminal Norte

Integra o setor Norte do Porto de Aveiro. Dispõe de 900 metros de cais com fundos à cota de -12 m (Z.H.) [1], 10 postos de acostagem, 8 armazéns a coberto, 277670 m^2 de terraplenos, ligação ferroviária à linha ferroviária nacional, 1 grua móvel com capacidade máxima de 35 toneladas e 6 gruas com capacidade máxima de 12 toneladas. Encontra-se vocacionado para a movimentação de carga geral e granéis sólidos, sendo o principal terminal polivalente do porto, que totaliza um tráfego anual de 1,8 milhões de toneladas. Tem como principais mercadorias movimentadas o cimento, cereais, pasta de papel, perfilados metálicos, aglomerados de madeira e argilas.

Terminal de Contentores e *Roll On Roll Off*

Com o intuito de alcançar novos mercados, a APA apostou em 2 novos segmentos: os contentores e carga *Roll on Roll Of* (RO-RO). Este terminal compreende um cais de 450 metros, fundos à cota de -12 m (Z.H.) e uma rampa para Tráfego Marítimo RO-RO. Devido aos 138000 m^2 de terraplenos disponíveis, é um terminal com um potencial de expansão elevado, que pode ser aproveitado para instalação de novos serviços.

Terminal Sul

A exploração comercial da operação portuária deste terminal encontra-se concessionada, em regime de serviço público, à empresa SOCARPOR - Sociedade de Cargas Portuárias (Aveiro), S.A.
Tal como o terminal Norte, é uma infraestrutura multiusos, como o terminal Norte, e oferece um cais de 400 metros, fundos à cota de -7 m (Z.H.) e cerca de 5 ha de terraplenos. Movimenta sobretudo produtos metalúrgicos, cimento, pasta de papel e produtos agroalimentares.

Terminal de Granéis Líquidos

As suas instalações encontram-se exploradas por diversas entidades privadas da indústria química que se dedicam à movimentação e armazenagem de produtos químicos, produtos

[1] Cotas Ortométricas referenciadas ao Zero Hidrográfico, em relação ao nível médio das águas do mar.

vitivinícolas e derivados de petróleo.

O terminal disponibiliza 6 pontes-cais, onde três delas com fundos à cota -12 m (Z.H.) e as restantes à cota -7 m (Z.H.).

Terminal de Granéis Sólidos

Este terminal está direcionado a dois segmentos de granéis: agro-alimentar e outros. Totaliza 750 m de cais com uma área de terraplenos de 151 mil m^2 devidamente equipados para instalação de serviços de valor acrescentado.

Porto de Pesca Costeira

Este terminal especializado dispõe de um conjunto de infra-estruturas destinadas à descarga, armazenagem e comercialização de pescado, para os comerciantes locais.

A lota e a fábrica de gelo encontram-se concessionadas à empresa Docapesca, Portos e Lotas, S.A.

Situado junto ao Porto de Pesca Costeira, existe o Porto de Abrigo. Este tem uma capacidade para 136 embarcações, possui um edifício de apoio e 72 armazéns de aprestos, destinados ao apoio das atividades dos utentes do terminal.

Porto de Pesca de Largo

Constituído por 17 pontes-cais com fundos de aproximadamente -7 m (Z.H.), este terminal serve os armadores de pesca do largo e as indústrias de processamento do pescado instaladas na Gafanha da Nazaré. Inclui um Terminal Especializado de Descarga de Pescado com 160 metros de comprimento, equipado com as infra-estruturas necessárias para suportar o funcionamento de uma unidade desta natureza.

ZALI - Zona de Atividades Logísticas e Industriais

Entre o terminal RO-RO e o de Granéis Sólidos está disponível uma reserva de terrenos, com o objetivo de se implementar atividades logísticas e industriais. Esta plataforma portuária tem como missão facilitar a implantação de operadores logísticos e empresas que tenham proveito na proximidade ao porto.

Plataforma Multimodal de Cacia

A plataforma multimodal de Cacia é a ligação ferroviária compatível com a bitola europeia e dista cerca de 8,8 km de distância dos terminais portuários de Aveiro.

2.1.4 Acessibilidades

Acessibilidade Marítima

Atualmente, o Porto de Aveiro recebe navios com cerca de 10,5 metros de calado e 150 metros de comprimento, podendo, em condições de navegação ideais, receber navios com um comprimento superior.

Durante o ano de 2013, foram realizadas dragagens para aprofundamento do canal de navegação principal do porto, no âmbito do projeto "Intervenção na Zona da Barra de Aveiro com Dragagem e Reforço do Cordão Dunar". Consequentemente, este conseguiu aumentar a capacidade de entrada de navios de maior dimensão, através da estabilização da barra de acesso marítimo à cota -12,5 m (Z.H.). Este investimento deveu-se à constante procura do porto de se tornar mais competitivo quer nacional quer internacionalmente.

Figura 2.3: Barra de Acesso ao Porto de Aveiro.

A entrada da barra do Porto de Aveiro (ver figura 2.3) situa-se a 1,5 milhas dos principais terminais do Setor Norte e a 4,5 milhas do Setor Sul, localizado no concelho de Aveiro.

A entrada no porto e no canal principal de navegação estão claramente assinaladas, não só pelo farol da barra, mas também por dois farolins localizados nas cabeças dos molhes norte e sul.

Acessibilidade Rodoviária

O acesso rodoviário ao Porto de Aveiro (ver figura 2.4) é constituído pelas autoestradas A1, A29, A25, A27, A17 e a ligação à IP3. O Porto de Aveiro está assim ligado às principais cidades portuguesas e ao centro de Espanha, o que se traduz numa excelente ligação terrestre à sua *hinterland*. Estas vias rodoviárias encontram-se num bom estado de

conservação, o que evita o congestionamento e facilita a ligação terrestre entre os diferentes terminais portuários.

Figura 2.4: Acesso Rodoviário ao Porto de Aveiro.

Acessibilidade Ferroviária

A construção da ligação ferroviária ao Porto de Aveiro iniciou-se em 2007 e contemplou três fases, até ao dia 27 de março de 2010, em que o ramal ferroviário foi finalmente inaugurado. Os fatores determinantes para a construção desta obra foram a importância que o Porto de Aveiro assume e a concentração de um considerável número de indústrias em território nacional. Este ramal visa novas ligações ferroviárias, através das quais faz-se a movimentação de mercadorias, que anteriormente se fazia através das vias marítima ou rodoviária. A ligação ferroviária permite também ao porto uma maior integração na Rede Transeuropeia de Transportes, em especial no corredor E-80, onde o porto possui uma localização privilegiada.

Figura 2.5: Ramal Ferroviário do Porto de Aveiro.

O ramal do Porto de Aveiro (ver figura 2.5) tem uma extensão de cerca de nove quilómetros entre o porto e a Plataforma Multimodal de Cacia. Com um feixe distribuidor de cinco ramificações, permite a ligação aos terminais portuários. No entanto, existe a possibilidade de, no futuro, se expandir o ramal para 9 feixes.

Aquando da sua construção, o seu principal objetivo era a maior captação de mercadorias líquidas e sólidas, de produtos agro-alimentares e de carga contentorizada. Apesar disso, atualmente, a ferrovia tem o cimento como prin-

cipal mercadoria movimentada. Desde a sua génese, a movimentação de mercadorias via ferrovia tem aumentado muito, contribuindo para o crescimento do porto.

2.1.5 Organograma

A APA, S.A. é organizada segundo o organograma resumido da figura 2.6.

O estágio esteve inserido no Gabinete de Estatística que se encontra incluído na Direção Financeira e de Desenvolvimento Organizacional.

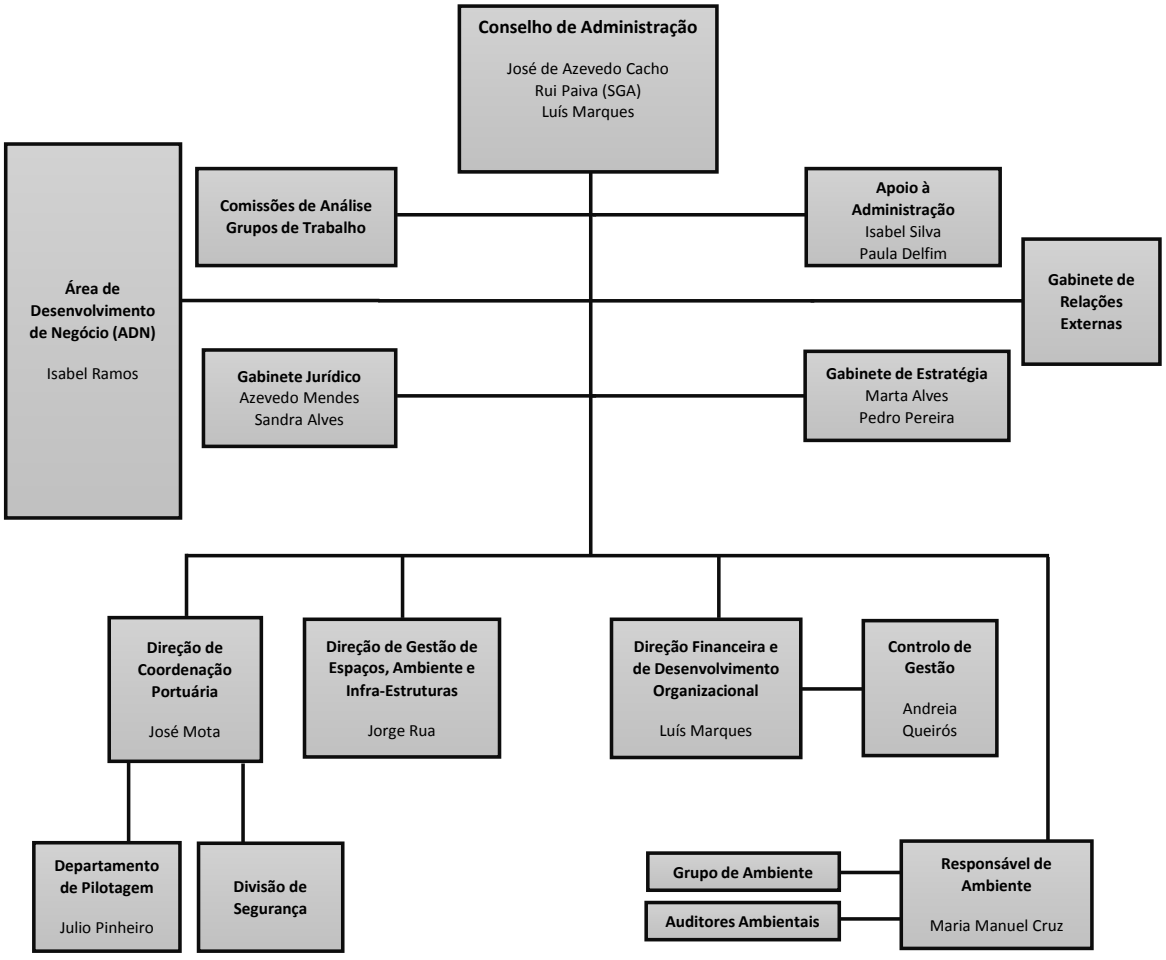

Figura 2.6: Organograma resumido da APA, S.A.

2.2 Porto da Figueira da Foz

A génese do Porto da Figueira da Foz data a 1166, mas só em novembro de 2008 é criada a Administração do Porto da Figueira da Foz, S.A., designada como sociedade anónima de capitais exclusivamente públicos, abreviadamente intitulada de APFF, S.A., com capital integralmente participado pela APA, S.A..

O Porto da Figueira da Foz (ver figura 2.7) dispõe de um terminal de Carga Geral, um terminal de Granéis, uma Doca de Recreio, uma Doca dos Bacalhoeiros, um Terminal de Receção de Produtos Betuminosos, um Porto de Pesca Costeira e um Portinho de Gala para Pequena Pesca.

O gabarito máximo dos navios com acesso marítimo ao Porto da Figueira da Foz é de cerca de 6 metros de calado e 120 metros de comprimento. Isto verifica-se, devido às características da barra de acesso. Relativamente aos acessos rodoviários, o porto da Figueira da Foz possui um conjunto de acessos rodoviários que podem ser equiparados a autoestradas.

Figura 2.7: Porto da Figueira da Foz.

2.3 Atividades Realizadas no Estágio

O estágio esteve inserido no Gabinete de Estatística, tendo sido orientado pelo Dr. Luís Sousa. Inicialmente, existiu uma fase de integração no porto, onde foram dados a conhecer as diversas áreas de trabalho e os diversos terminais portuários.

As tarefas realizadas durante o tempo de estágio centraram-se na estatística portuária: uma estatística bastante descritiva. As principais variáveis de interesse ao porto são: o

número de navios, a quantidade e o tipo de mercadoria movimentada, o tipo de operação (carga ou descarga), o tempo de estadia dos navios no porto, os agentes, os operadores, entre outros indicadores que tenham relevância. Com base nisto, foram realizados vários estudos estatísticos, a pedido dos quadros superiores da empresa.

Para a realização dos estudos estatísticos foi utilizado o *software* Microsoft Excel, que disponibiliza as ferramentas necessárias para a obtenção da estatística exigida pelo porto. Para a obtenção dos dados, foi disponibilizado o acesso aos programas SIGAPA e SIGFOZ, onde se pode encontrar o módulo de estatística, faturação, tarifário e gestão dominial da APA, S.A. e APFF, S.A., respetivamente, a JUP - Janela Única Portuária, e folhas de cálculo elaboradas pelos colaboradores no cais de exploração.

Mensalmente, foram atualizadas as áreas de estatística dos portais (sites) do Porto de Aveiro e Porto da Figueira da Foz, preenchendo *templates* de tabelas disponibilizados pelo orientador no porto. Nestas tabelas,são pedidas as seguintes informações: a quantidade de mercadoria em relação a cada tipo de carga, os valores acumulados ao longo do ano, o número de navios e algumas características dos mesmos, nomeadamente, a arqueação bruta total e o comprimento total dos navios. São também elaborados os *reportings* legais destinados ao Instituto Nacional de Estatística (INE) e ao Instituto Portuário e dos Transportes Marítimos (IPTM), tanto para o Porto de Aveiro, como para a Figueira da Foz.

Trimestralmente, é habitualmente solicitada informação mais detalhada a nível estatístico, de forma a analisar o trimestre e compará-lo com outros anos. Os estudos abrangem sempre o Porto de Aveiro e o Porto da Figueira da Foz.

Foram realizadas também tarefas não periódicas como:

- Apoio na elaboração de artigos destinados a divulgação na empresa e nos portais;
- Participação no fórum empresarial do mar, representando o Porto de Aveiro, com o acompanhamento do orientador, onde foi estabelecido um contacto mais próximo com pessoas do setor portuário;
- Visitas a portos de referência nacional, como o Porto de Sines e Porto de Setúbal, onde existiu a oportunidade de conhecer os gabinetes estatísticos dos portos referidos, conseguindo, posteriormente, retirar pontos de aprendizagem para o Porto de Aveiro e Porto da Figueira da Foz;

- Verificação e validação da integridade e qualidade dos dados da base de dados de estatística do Porto de Aveiro referentes ao ano de 2013, onde foram analisados os registos respeitantes a cada processo de navio, tendo em consideração a codificação das mercadorias, navios e tipo de operação, as quantidades de mercadorias movimentadas, entre outros.

Capítulo 3

Enquadramento Teórico

3.1 Modelos de Regressão Múltipla

A regressão linear múltipla é uma técnica multivariada que tem como propósito estabelecer uma relação entre uma variável dependente (ou resposta) e duas ou mais variáveis explicativas. Através da regressão linear múltipla, pretende-se estudar de que forma as variáveis explicativas são significativas para justificar a variabilidade da variável resposta.

3.1.1 Caracterização do Modelo

Considere-se uma amostra da forma $(x_{i1}, x_{i2}, ..., x_{ip}, y_i)$, $i = 1, ..., n$, ou seja, uma amostra em que cada uma das n componentes é constituída pelos valores de p regressores e da correspondente variável resposta. Os regressores designam-se também por variáveis explicativas ou variáveis controladoras. A equação que caracteriza a relação entre as observações de y e as observações das variáveis explicativas é

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \quad (i = 1, ..., n), \tag{3.1}$$

onde $x_1, x_2, ..., x_p$ são variáveis explicativas determinísticas, $\beta_0, \beta_1, ..., \beta_p$ são coeficientes da equação de regressão e ϵ_i representa os erros do modelo, sob as seguintes hipóteses: ϵ_i não correlacionados, com $V(\epsilon_i) = \sigma^2$ e $E(\epsilon_i) = 0$; usualmente consideram-se normalmente distribuídos.

Em notação matricial, a representação do modelo torna-se mais compacta. Para reescrever o modelo na forma matricial, definimos um vetor $(n \times 1)$ de todas as observações da

variável resposta

$$y = [y_1, y_2, ..., y_n]^T,$$

um vetor $(n \times (p+1))$ dos coeficientes de regressão

$$\beta = [\beta_0, \beta_1, \beta_2, ..., \beta_p]^T,$$

um vetor $(n \times 1)$ dos erros do modelo

$$\epsilon = [\epsilon_1, \epsilon_2, ..., \epsilon_n]^T,$$

e uma matriz $(n \times (p+1))$ de observações dos regressores

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

Tendo em conta as definições acima, chegamos à seguinte notação matricial do modelo de regressão múltipla:

$$y = X\beta + \epsilon. \tag{3.2}$$

Subjacente a este modelo, é necessário assumir hipóteses quanto à distribuição dos erros:

1. $E[\epsilon] = 0$;
2. $Cov[\epsilon] = \sigma^2 I_n$.

Então, a equação de regressão é

$$E[y] = X\beta \tag{3.3}$$

e assumindo a distribuição Normal, ter-se-á que $y \sim N(X\beta, \sigma^2 I_n)$.

3.1.2 Estimação dos Parâmetros

Nesta secção, trataremos de estimar os parâmetros de regressão. Para tal, utiliza-se o método dos mínimos quadrados para estimar $\hat{\beta}$. Assim, o nosso objetivo é encontrar o

vetor $\hat{\beta}$, que minimiza

$$\begin{aligned} S &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta} \end{aligned}$$

Através da derivação de S em ordem a $\hat{\beta}$ obtém-se

$$\frac{\delta S}{\delta \hat{\beta}} = 0 - 2X^T y + 2X^T X\hat{\beta}.$$

Pelo que se obtém o estimador definido por

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{3.4}$$

Esta solução é válida, desde que a matriz $X^T X$ seja não singular.

Pode-se notar de imediato que este estimador possui as seguintes propriedades:

$$E[\hat{\beta}] = \beta$$

$$\Sigma_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$$

Note-se que o estimador é centrado, os elementos diagonais de $\Sigma_{\hat{\beta}}$ descrevem a variância de cada componente $\hat{\beta}_j$ e os elementos (i, j) com $i \neq j$ de $\Sigma_{\hat{\beta}}$ representam as covariâncias entre as componentes $\hat{\beta}_i$ e $\hat{\beta}_j$ do estimador $\hat{\beta}$.

Para utilizar várias fórmulas anteriores é necessário estimar σ^2 que geralmente é desconhecido. É possível estimar σ^2 a partir dos resíduos e_i. Através da soma dos quadrados dos resíduos obtém-se

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} {e_i}^2 \\ &= e^T e. \end{aligned}$$

Substituindo $e = y - X\hat{\beta}$, temos

$$\begin{aligned} SS_{Res} &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta} \\ &= y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}. \end{aligned}$$

Sendo que $X^T X\hat{\beta} = X^T y$, a equação anterior reescreve-se da seguinte forma

$$SS_{Res} = y^T y - \hat{\beta}^T X^T y.$$

A média do quadrado dos resíduos é dada por

$$MS_{Res} = \frac{SS_{Res}}{n-p},$$

onde $n - p$ são os graus de liberdade associados à soma do quadrado dos resíduos.

Tendo em conta que o valor esperado da média do quadrado dos resíduos é σ^2, então um estimador não enviesado de σ^2 é:

$$\hat{\sigma}^2 = MS_{Res}.$$

3.1.3 ANOVA da Regressão

Teste de significância do modelo

A questão que agora se coloca é se existe ou não algum benefício em relacionar a variável resposta y com as variáveis explicativas x_j para $j = 1, \ldots, p$. Formalmente, pode-se testar a significância da regressão através do seguinte teste de hipóteses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad vs. \quad H_1 : \beta_j \neq 0 \text{ para pelo menos um valor de j.} \tag{3.5}$$

O teste (3.5) pretende determinar se de facto existe uma relação linear entre a variável resposta y e qualquer um dos regressores x_j. A rejeição da hipótese nula significa que pelo menos um dos regressores x_j contribui significativamente para o modelo.

O procedimento do teste é uma generalização da análise da variância com base na tabela ANOVA. A soma quadrática total SS_T é decomposta pela soma dos quadrados dos

resíduos, SS_{Res}, e a soma dos quadrados da regressão, SS_R. Isto é,

$$\begin{aligned} & SS_T = SS_R + SS_{Res} \\ \iff \quad & \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \end{aligned}$$

Recorrendo à notação matricial, verifica-se que o cálculo das somas quadráticas pode também ser efetuado da seguinte forma: $SS_{Res} = \hat{\beta} X^T y - n\bar{y}^2$ e $SS_R = y^T y - \hat{\beta} X^T y$. Verifica-se que o número de graus de liberdade se apresenta repartido do seguinte modo:

$$\begin{aligned} & df(SS_T) = df(SS_{Res}) + df(SS_R) \\ \iff \quad & n - 1 = (n - p - 1) + p. \end{aligned}$$

Através da divisão das somas quadráticas pelos respetivos graus de liberdade obtêm-se as correspondentes "médias quadráticas":

$$MS_{Res} = \frac{SS_{Res}}{n - p - 1}, \quad MS_R = \frac{SS_R}{p}.$$

Podem-se sintetizar todas estas características na seguinte tabela ANOVA:

Tabela 3.1: Tabela ANOVA.

Fonte de Variação	**Soma de Quadrados**	**Graus de liberdade**	**Quadrados Médios**	**Estatística**
Regressão	SS_R	p	MS_R	$F^* = MS_R/MS_{Res}$
Resíduos	SS_{Res}	$n-(p+1)$	MS_{Res}	
Total	SS_T	$n-1$		

Se a escolha do modelo for adequada, a estatística de teste F^* deve ser muito superior a 1. Isto porque, a maior parte da variação em y deve ter como causa a equação de regressão, ou seja, SS_R deverá ser muito maior do que SS_{Res}.

Supondo que os erros do modelo têm distribuição Normal, sabe-se que SS_R e SS_{Res} são independentes, que $\frac{SS_R}{\sigma^2}$ e $\frac{SS_{Res}}{\sigma^2}$ seguem distribuições qui-quadrado (χ^2) com p e $n-p-1$ graus de liberdade, respetivamente, e que a estatística F^* segue uma distribuição de Fisher com $(p, n-p-1)$ graus de liberdade.

O teste de hipóteses (3.5) tem como estatística de teste F^*. A hipótese nula é rejeitada

quando a estatística de teste apresenta valores elevados, o que significa que alguns regressores contribuem significativamente para explicar a variabilidade existente na variável resposta.

3.1.4 Inferências sobre Coeficientes e Predições

Atendendo ao facto de estarmos a pressupor a normalidade para os erros, o estimador $\hat{\beta}$ de mínimos quadrados tem distribuição normal multivariada $N(\beta, \sigma^2(X^TX)^{-1})$. Isto implica que a distribuição marginal de $\hat{\beta}_j$, estimador do coeficiente de regressão β_j, segue uma distribuição normal com média β_j e variância $\sigma^2 c_{jj}$; ou seja,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{c_{jj}}} \sim N(0,1), \text{onde } c_{jj} \text{ é o } j\text{-ésimo elemento da diagonal da matriz } (X^TX)^{-1}.$$

Estimando a variância dos erros do modelo, conclui-se que

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{(n-p-1)}, \tag{3.6}$$

onde $t_{(n-p-1)}$ representa a distribuição t-Student com $n-p-1$ graus de liberdade e sendo $s(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$ o erro padrão de cada estimador individual $\hat{\beta}_j$, para $j = 1, \ldots, p$.

O resultado (3.6) permite-nos determinar um intervalo de confiança para β_j com facilidade, concluindo que é dado por

$$\hat{\beta}_j - t_{\frac{\alpha}{2},n-p-1} \times s(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2},n-p-1} \times s(\hat{\beta}_j),$$

com um grau de confiança $(1-\alpha)\times 100\%$, onde $t_{\frac{\alpha}{2},n-p-1}$ representa o quantil de ordem $\frac{\alpha}{2}$ de uma distribuição t-Student com $n-p-1$ graus de liberdade.

A partir do resultado (3.6) pode-se também averiguar a significância dos regressores para o modelo. Cada β_j traduz a contribuição do regressor x_j para explicar a variação em y. Se o valor de β_j for quase nulo conclui-se que não é uma contribuição importante.
Consequentemente, pode-se testar a significância de cada regressor, através do teste de hipóteses

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0, j = 1, \ldots, p. \tag{3.7}$$

Associado ao teste de hipóteses (3.7) temos a seguinte estatística de teste

$$T_j = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}. \tag{3.8}$$

Para um nível de significância α, rejeita-se a hipótese nula, H_0 em (3.7), quando o valor observado da estatística de teste (3.8) for significativamente diferente de 0, isto é, encontra-se fora do intervalo $(-t_{\frac{\alpha}{2},n-p-1}, t_{\frac{\alpha}{2},n-p-1})$. Se a hipótese nula do teste for rejeitada pode-se concluir que o regressor x_j contribui significativamente para explicar a variação em y estando os restantes regressores presentes no modelo. É necessário prudência na interpretação do resultado, pois está-se a testar a significância de x_j com a presença das restantes variáveis no modelo. Assim se $\hat{\beta}_j$ não for significativamente diferente de zero significa que x_j não contribui de uma forma significativa para a variação de y após as contribuições de todas as outras variáveis explicativas terem sido consideradas. Portanto, é possível que a variável x_j seja significativa se alguma das outras variáveis explicativas forem retiradas do modelo. Este facto é importante quando se selecionam as variáveis explicativas que devem constar no modelo.

Quando se pretende predizer um valor futuro para y, assumindo que os valores das variáveis explicativas x_j para $j = 1, \ldots, p$ são conhecidos, importa distinguir dois casos: predizer o valor médio das respostas, $\hat{y}_{n+1}$, ou prever um valor individual de resposta $y_{(n+1)}$.

Considere-se $x_{(n+1)}$ o futuro valor para o qual se pretende predizer a resposta. É evidente que as estimativas pontuais $y_{(n+1)}$ e $\hat{y}_{n+1}$ são dadas por

$$y_{(n+1)} = \hat{y}_{(n+1)} = x_{(n+1)}\hat{\beta}.$$

Facilmente se verifica então que os erros padrão são, respetivamente,

$$s(\hat{y}_{(n+1)}) = \hat{\sigma}\sqrt{x^T{}_{(n+1)}(X^TX)^{-1}x_{(n+1)}}, \text{para o valor médio da resposta,}$$
$$s(y_{(n+1)}) = \hat{\sigma}\sqrt{1 + x^T{}_{(n+1)}(X^TX)^{-1}x_{(n+1)}}, \text{para o valor individual da resposta,}$$

onde σ^2 é estimado a partir da média quadrática dos resíduos. Pode-se notar que a variabilidade em $\hat{y}_{(n+1)}$ surge unicamente de flutuações de amostragem em $\hat{\beta}$, enquanto que a variabilidade em $y_{(n+1)}$ tem adicionalmente a variabilidade σ^2 causada pela variância dos valores individuais sobre a resposta média.

Conclui-se que as estimativas pontuais irão coincidir claramente para os dois casos, porém o erro padrão do estimador altera-se, sendo que, é maior aquando da predição de valores individuais da resposta.

Podem-se ainda determinar intervalos de confiança para observações futuras, tendo em consideração os erros padrões anteriormente apresentados, supondo a normalidade dos erros do modelo e recorrendo à distribuição $t_{(n-p-1)}$.

3.1.5 Avaliação da Qualidade e Significado da Regressão

Métodos Gráficos

No caso da Análise de Regressão Linear Simples as análises dos gráficos, de dispersão y_i contra x_i, são muito úteis. É bem sabido que se os dados forem bem modelados pelo modelo de Regressão Linear Simples os pontos devem dispor-se aproximadamente sobre uma reta. No entanto, no caso de termos vários regressores a situação complica-se e a análise de cada gráfico de dispersão considerando um regressor isoladamente não se revela tão eficaz. Os gráficos de y_i contra cada regressor x_i só serão informativos se a correlação entre os regressores for baixa.

Ryan,(1997) apresenta exemplos vários de sugestões claras de perfeitos ajustamentos lineares, e de situações contrárias, onde a simples análise do gráfico de dispersão pode indicar linearidade e de facto, ela não existir.

Coeficiente de Correlação Parciais

Além de se quantificar a correlação entre y e o conjunto de todas as variáveis $x_1, \ldots, x_p$, há frequentemente interesse em avaliar a correlação de y com cada uma destas variáveis, tomadas individualmente. No entanto, quando estão envolvidas diversas variáveis independentes, o grau de correlação simples entre y e cada um dos regressores não é uma medida suficiente para expressar o seu grau de relacionamento linear.

É de toda a relevância considerar medidas de correlação parciais, que permitem definir correlações entre duas variáveis quaisquer, quando se anulam os efeitos induzidos pelas restantes. Sejam X, Y e Z três variáveis aleatórias e ρ_{XY}, ρ_{XZ} e ρ_{YZ} os coeficientes de correlação amostral entre as variáveis. Assim, o coeficiente de correlação parcial entre X e

Y quando Z se mantém constante é dado por:

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ} \cdot \rho_{YZ}}{\sqrt{(1 - \rho^2{}_{XZ})(1 - \rho^2{}_{YZ})}}.$$

Por exemplo, se existirem 3 variáveis, $\rho_{xy|z}$ traduz a correlação parcial entre x e y quando z é mantido constante, e com

$$\rho_{xxy|z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{(1 - \rho^2{}_{xz})(1 - \rho^2{}_{yz})}}$$

e ρ_{xy} representa a correlação amostral entre x e y.

A análise das correlações parciais tem particular interesse no contexto da análise de correlação múltipla porque:

- permite calcular a correlação de uma variável independente e a variável dependente quando já existem outras variáveis independentes na regressão;

- representa o efeito preditivo do incremento provocado pela variável independente que não é explicado pelas variáveis independentes que já estão incluídas na regressão;

- tem como grande utilidade o de identificar as variáveis independentes com maior poder preditivo incremental.

Coeficiente de Determinação

O Coeficiente de Determinação definido por

$$R^2 = \frac{SS_R}{SS_T}$$

é um indicador que permite avaliar a qualidade da regressão sem assumir a normalidade dos erros. Através deste coeficiente pode-se estudar a proporção da variação da variável resposta y que é explicada pela equação de regressão considerando os p regressores. O coeficiente toma valores entre 0 e 1. Contudo, um grande valor de R^2 não significa obrigatoriamente que o modelo de regressão seja um bom ajustamento, isto porque a adição de uma variável aumenta sempre o valor deste coeficiente, sem ter em consideração se a variável que se adiciona é ou não significativa para o modelo. Por este motivo R^2 não será um bom indicador do grau de ajustamento do modelo. Sendo assim, é preferível utilizar o

Coeficiente de Determinação Ajustado que se define por

$$R^2{}_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{MS_{Res}}{MS_T}.$$

Este coeficiente dá uma melhor ideia da proporção de variação de y explicada pelo modelo de regressão pois apenas aumenta quando a introdução de novos regressores for vantajosa. Quando a diferença entre R^2 e $R^2{}_{adj}$ é acentuada, significa que existe uma boa possibilidade de que tenham sido incluídos no modelo regressores estatisticamente não significativos.

Variância dos Erros

Deve-se comparar a estimativa do desvio padrão dos erros $\hat{\sigma} = \sqrt{MS_{Res}}$ com a estimativa, s, do desvio padrão da amostra das observações da variável dependente. Se a estimativa $\hat{\sigma}$ não for significativamente inferior a s, significa que para predizer a variável dependente y, basta considerar a média $\bar{y}$ das observações, pois o modelo de regressão não traz qualquer vantagem em termos de predição de futuras observações. (Hall et al, 2006)

3.1.6 Validação dos Pressupostos da Regressão

Análise de Resíduos

Em estudos de regressão devem-se examinar os resíduos de forma a verificar os pressupostos do modelo e identificar valores atípicos que possam estar presentes. Para a análise de regressão assumimos que os erros ϵ_i seguem uma distribuição Normal com média nula e variância σ^2 e que são independentes, uma vez que são não correlacionados. Assim, os resíduos e_i (estimativa dos erros ϵ_i) devem refletir as propriedades dos erros: serem normais, independentes e terem variância constante.

Tendo sido obtido o estimador dos mínimos quadrados $\hat{\beta}$, expresso em (3.4), tem-se que

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X(X^TX)^{-1}X^Ty.\end{aligned}$$

Assim, pode-se reescrever da seguinte forma

$$\hat{y} = Hy,$$

onde, $H = X(X^TX)^{-1}X^T$. Esta matriz é conhecida como a matriz chapéu e, através dela, podem-se exprimir os resíduos matricialmente, uma vez que

$$\begin{aligned} e &= y - \hat{y} \\ &= y - X\hat{\beta} \\ &= y - Hy \\ &= (I - H)y. \end{aligned}$$

Normalidade dos Erros

A análise gráfica dos resíduos é uma maneira de investigar preliminarmente a adequação do ajustamento do modelo.

Um método muito simples de verificar a hipótese de normalidade é a construção de um *PP-plot* ou de um *QQ-plot*. No caso do *PP-plot*, isto é, do gráfico de probabilidade normal (*Normal Probability Plot*), pretende-se representar a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade acumulada dos erros. O *QQ-plot*, isto é, o gráfico de quantis, compara os quantis empíricos (amostrais) com os quantis da distribuição Normal. Se os erros possuírem uma distribuição normal, todos os pontos do gráfico *PP-plot* ou *QQ-plot* devem posicionar-se mais ou menos sobre uma reta. Se a distribuição dos erros não for bem modelada por uma normal, quer o *QQ-plot* ou o *PP-plot* apresentam curvaturas numa ou mais zonas do gráfico.

Analiticamente, podem-se ainda utilizar testes de ajustamento como o teste de Kolmogorov - Smirnov e o teste de Shapiro - Wilk para verificar a normalidade dos resíduos, apresentados seguidamente:

1. **Teste de Kolmogorov-Smirnov**

   O teste de Kolmogorov-Smirnov avalia se uma amostra foi retirada de uma população com uma distribuição de probabilidade específica. O método tem por base a distância entre a função de distribuição empírica $F_n(x)$ e a função de distribuição fixada na hipótese nula $F_0(x)$, que se pretende que seja o mais pequena possível. Assim, pretende-se testar:

   $$H_0 : F_n(x) = F_0(x) \qquad versus \qquad H_1 : F_n(x) \neq F_0(x).$$

O teste de Kolmogorov-Smirnov tem como estatística de teste:

$$D = \sup_{x \in \mathbb{R}} \mid F_n(x) - F_0(x) \mid,$$

onde, sup designa supremo. A estatística de teste D representa a maior distância entre a função de distribuição empírica e a função de distribuição em teste.

Rejeita-se a hipótese nula para valores muito elevados da estatística de teste. Portanto, a região crítica apropriada, ao nível de significância α é

$$RC_\alpha = \{d : d > c_\alpha\},$$

onde, os valores críticos c_α se encontram tabelados. A interpretação do resultado pode ser também efetuada a partir do *p-value* do teste. Se *p-value* $\leq \alpha$ então rejeita-se a hipótese de ajustamento.

É importante ter em atenção que o teste de Kolmogorov-Smirnov só deve ser aplicado a amostras de dimensão não muito reduzida senão a potência do teste tende a ser muito baixa.

No caso em que se desconhecem os parâmetros da distribuição é necessário estimá-los recorrendo às observações amostrais. Esta estimação implica uma redução na potência do teste. Com o intuito de combater esta falha, *Lilliefors* desenvolveu uma correção ao teste de Kolmogorov-Smirnov para o caso da distribuição em teste ser a Normal. Este teste deve ser aplicado a amostras de dimensão elevada.

2. **Teste de Shapiro-Wilk**

   O teste de Shapiro-Wilk testa a normalidade de uma dada amostra ou população. É o teste mais indicado quando o tamanho da amostra é reduzido, isto é, menor que 50.
   Pode-se então formular as hipóteses do teste da seguinte forma:

$$H_0 : X \text{ provém de uma distribuição Normal} \quad versus \quad H_1 : \text{caso contrário.}$$

   O teste de Shapiro-Wilk é baseado na estatística W dada por:

$$W = \frac{\sum_{i=1}^{n} a_i {X_i}^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

onde, para $i = 1, \cdots, n$, X_i são os valores da variável X, $\bar{X}$ representa a média de X e a_i são as constantes geradas através da média, variância, e covariânvia de n ordens com a distribuição Normal. Estes valores encontram-se tabelados.

Para valores pequenos de W rejeita-se a hipótese da normalidade. Se $W < W_\alpha$ então rejeita-se a hipótese nula para um dado nível de significância α e para valores tabelados de $W_{\alpha,n}$. Para um nível de significância α, se o *p-value* $\leq \alpha$ então rejeita-se a hipótese nula.

Aleatoriedade dos Erros e Homocedasticidade do Modelo

Para averiguar a independência dos erros e o caráter constante da respetiva variância, constroem-se gráficos de resíduos e_i contra os valores preditos $\hat{y}_i$. Se o gráfico obtido corresponder a uma mancha de pontos aleatórios com o mesmo tipo de dispersão em torno do eixo das abcissas, isto é, contidos aleatoriamente numa faixa horizontal, não existem motivos para pôr em causa os referidos pressupostos. Tipicamente, se, apesar de não se rejeitar a normalidade, os gráficos dos resíduos deixarem de parecer uma mancha aleatória passando a apresentar tendências, tal sugere a rejeição da homocedasticidade e independência dos erros. Em particular, o padrão de funil dos resíduos implica que a variância dos erros não é constante.

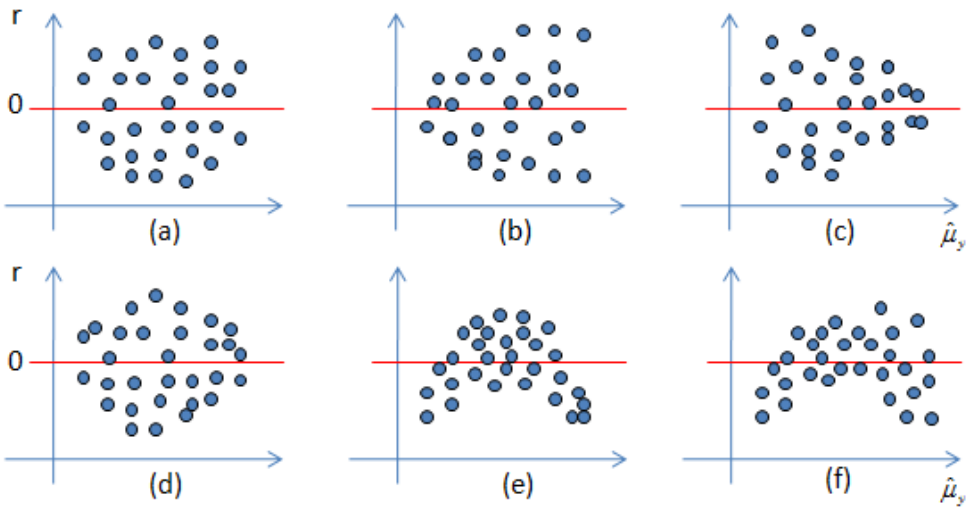

Figura 3.1: Padrões para os gráficos de resíduos: (a) aleatoriedade; (b) e (c) funil; (d) duplo arco; (e) e (f) não linear.

Veja-se em particular cada uma das situações:

1. **Independência**

Teste de Durbin-Watson

O teste de Durbin-Watson tem como objetivo detetar a presença de autocorrelação (dependência) nos resíduos numa análise de regressão.

Testamos a presença de autocorrelação por meio das hipóteses

$$H_0 : \rho = 0 \qquad versus \qquad H1 : \rho > 0.$$

A estatística de teste é dada por

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2},$$

onde $e_i = y_i - \hat{y}_i$ e y_i e $\hat{y}_i$ são, respetivamente, os valores observados e previstos da variável resposta para cada i. O valor de d torna-se menor com o aumento das correlações. Os valores críticos superiores e inferiores, d_U e d_L estão tabelados para diferentes valores de p e n (ver apêndice A). Portanto:

i. se $d < d_L$ rejeita-se H_0;

ii. se $d > d_U$ não se rejeita H_0;

iii. se $d_L < d < d_U$ o teste é inconclusivo.

No caso de se concluir que os erros são correlacionados, usa-se o método de mínimos quadrados generalizados. Considere-se $var(\epsilon_i) = \sigma^2\Omega$, onde Ω é uma matriz simétrica, definida positiva, que representa as variâncias e covariâncias dos erros. No caso de Ω ser conhecida, temos que o estimador do método dos mínimos quadrados generalizados para β é dado por

$$b_{MQG} = \left(X^T\Omega^{-1}X\right)^{-1}X^T\Omega^{-1}y.$$

Se a matriz Ω for desconhecida é necessário estimá-la. Considerando $\hat{\Omega}$ uma estimativa de Ω, então $b_{EMQG} = \left(X^T\hat{\Omega}^{-1}X\right)^{-1}X^T\hat{\Omega}^{-1}y$ é uma estimativa do estimador do método dos mínimos quadrados generalizados de β. Geralmente, as propriedades destas estimativas são difíceis de obter para pequenas amostras, exceto quando especificado no modelo métodos de Monte Carlo. No entanto, são fáceis de obter propriedades assintóticas.

2. **Homocedasticidade**

A primeira forma de detetar a existência de heterocedasticidade é, como se viu, através da análise gráfica dos resíduos, isto é, construir um gráfico de dispersão entre os resíduos e os valores ajustados da variável resposta. Para o diagnóstico da heterocedasticidade é necessário verificar se existe alguma tendência no gráfico; se os pontos mostrarem um padrão consistente, como por exemplo um "funil", significa que existem indícios de heterocedasticidade. Neste caso, é pertinente efetuar testes formais para a deteção da heterocedasticidade.

Em seguida serão abordados dois testes: o teste de Breusch-Pagan e o teste de Koenker.

Teste de Breusch-Pagan

O teste de Breusch-Pagan é bastante utilizado para testar a hipótese nula de que as variâncias dos erros são iguais (homocedasticidade) versus a hipótese alternativa de que alguma das variâncias seja diferente. É indicado para grandes amostras e quando a suposição de normalidade dos erros é assumida.

Primeiramente, ajusta-se o modelo de regressão linear múltipla de forma a obter os resíduos $\hat{e_i}$ e os valores ajustados $\hat{y_i}$, para $i = 1, \cdots, n$. Os resíduos da regressão original são usados para determinar os quadrados dos resíduos padronizados:

$$\hat{g_i} = \frac{\hat{e_i}^2}{\frac{\sum \hat{e_i}^2}{n}} \qquad i = 1, \cdots, n.$$

Usando $\hat{g_i}$ como a variável dependente, faz-se a regressão

$$\hat{g_i} = \hat{c_0} + \hat{c_1} Z_{i1} + \hat{c_2} Z_{i2} + \cdots + \hat{c_p} Z_{ip},$$

onde Z_{ip} são as variáveis que se pensa ser a origem da heterocedasticidade. Usando a soma dos quadrados da regressão (SS_{Res}) da segunda regressão é possível calcular a estatística de teste:

$$BP = \frac{SS_{Res}}{2}$$

que segue uma distribuição qui-quadrado com $p - 1$ graus de liberdade. Rejeita-se a hipótese nula, isto é a hipótese de homocedasticidade, se $BP > {\chi^2}_{\alpha,p-1}$ para um

dado valor de significância α. Resumidamente, se não existe heteroscedasticidade, é de se esperar que o quadrado dos resíduos não aumente ou diminua com o aumento do valor predito, $\hat{y}_i$ e assim, a estatística de teste deveria ser insignificante, isto é, levando à rejeição da igualdade das variâncias para valores elevados da estatística de teste BP. Se o $p - value \leq \alpha$ então rejeita-se a hipótese nula.

Teste de Koenker

Uma vez que o teste de Breusch-Pagan é sensível ao pressuposto da normalidade, Koenker (1981) propõe o seguinte teste baseado num estimador mais robusto.

Consideram-se os resíduos e_i, para $i = 1, \cdots, n$. Se a hipótese nula é verdadeira, isto é, se existe homocedasticidade nos erros, então temos que:

$$\hat{\sigma}^2 = \frac{1}{n} \sum {e_i}^2$$

se pode tomar para estimar a variância. Sejam $A = \sum \frac{{e_i}^2 - \hat{\sigma}^2}{n}$ e $\tilde{y} = \sum \frac{\hat{y}_i}{n}$; então a estatística de teste:

$$V = \frac{\{\sum {e_i}^2 (\hat{y}_i - \tilde{y})\}^2}{A \sum (\hat{y}_i - \tilde{y})^2},$$

segue, aproximadamente, uma distribuição qui-quadrado com um grau de liberdade quando a hipótese nula de homocedasticidade é verdadeira. Para um nível de significância α, rejeita-se a hipótese nula se $V \geq c$, onde c é o quantil $1 - \alpha$ da distribuição de qui-quadrado com um grau de liberdade. Estes valores encontram-se tabelados. Se o $p - value \leq \alpha$ então rejeita-se a hipótese nula.

Caso exista heterocedasticidade são possíveis duas ações corretivas para tornar as variâncias aproximadamente iguais. Uma delas é efetuar transformações na variável resposta de forma a estabilizar a variância. Quando se examina o diagrama de dispersão, o padrão mais comum é a distribuição em forma de cone. Se o cone abre à direita, considera-se habitualmente a inversa; se o cone abre à esquerda, considera-se a raiz quadrada. Outra ação corretiva envolve a ponderação da regressão através do método dos mínimos quadrados ponderados. O método dos mínimos quadrados ponderados é uma extensão do método dos mínimos quadrados clássico. Com a finalidade de estabilização das variâncias, o método dos mínimos quadrados ponderados na estimação dos parâmetros é mais adequado por fornecer estimadores não enviesados e de variância mínima.

Suponhamos que $var(\epsilon_i) = {\sigma_i}^2 = {c_i}^2\sigma^2$, onde ${c_i}^2$ são constantes conhecidas. Então, a variância constante também pode ser alcançada dividindo cada um dos lados da equação do modelo de regressão,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \cdots, n,$$

por c_i, isto é, considerar

$$\frac{y_i}{c_i} = \frac{\beta_0}{c_i} + \cdots + \frac{\beta_p x_{ip}}{c_i} + \frac{\epsilon_i}{c_i}, \quad i = 1, \cdots, n. \tag{3.9}$$

É claro que o modelo (3.9) é homocedástico. Cada $w_i = (c_i)^{-2}$ é denominado de peso. A estimação dos coeficientes β a partir do modelo (3.9) é feita através do método dos mínimos quadrados ponderados, e processa-se da forma habitual.

Para o caso das constantes ${c_i}^2$ serem desconhecidas é necessário definir Ω como uma matriz diagonal com os elementos da diagonal ${c_1}^2, \ldots, {c_n}^2$. Aplicando o método dos mínimos quadrados ponderados, o modelo original $y = X\beta + \epsilon$, com $E(\epsilon) = 0$ e $cov(\epsilon) = \sigma^2\Omega$, é transformado no modelo $y^{(\Omega)} = X^{(\Omega)}\beta + \epsilon^{(\Omega)}$, onde $y^{(\Omega)} = Cy$, $X^{(\Omega)} = CX$ e $\epsilon^{(\Omega)} = C\epsilon$ e C é uma matriz diagonal com elementos não nulos ${c_1}^{-1}, \ldots, {c_n}^{-1}$. Como $C\Omega C^T = I$, temos que $cov(\epsilon^{(\Omega)}) = \sigma^2 I$. Assim, pode-se prosseguir da forma habitual. Tendo em consideração que $CC^T = C^T C = \Omega^{-1}$, pode-se reescrever a estimativa de β da seguinte forma:

$$\begin{aligned} b_{MQP} &= \left(X^{(\Omega)^T} X^{(\Omega)}\right)^{-1} X^{(\Omega)^T} y^{(\Omega)} \\ &= \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} y. \end{aligned}$$

3.1.7 Seleção de Variáveis

A questão que se coloca nesta secção é se todos os regressores são importantes para a construção do modelo. Se um ou mais regressores não contribuírem significativamente para o modelo poderá fazer sentido não os considerar. Existem diversos procedimentos para encontrar um subconjunto apropriado de variáveis explicativas para o modelo. No entanto, é importante ter em consideração que o aumento do número de regressores diminui também os graus de liberdade de algumas estatísticas e como tal aumenta a incerteza de alguns resultados.

Por omissão, ao utilizar o *software* estatístico SPSS, o método de seleção das variáveis explicativas para o modelo é o método que introduz todas as variáveis, normalmente designado por "Enter". Mas existem outros métodos como iremos ver.

Seleção *Backward*

Este método começa com um modelo que inclui todos os p regressores. Em seguida, é calculada a estatística $F_{parcial}$ para cada regressor como se este fosse o último a entrar no modelo. A estatística denominada $F_{parcial}$ é calculada da seguinte forma

$$F_{parcial} = \frac{SS_R(x_k|x_1, \ldots, x_{k-1})}{MS_{Res}}$$

onde, $MS_{Res} = \frac{SS_{Res}}{n-p-1}$ representa o quadrado médio dos erros do modelo.

De seguida, o menor dos valores da estatística $F_{parcial}$ é comparado com o valor crítico da estatística já pré-estabelecido. Se o menor valor de $F_{parcial}$ for menor que o valor crítico então o regressor correspondente é removido do modelo. Agora, fica-se com um modelo de regressão com $p-1$ regressores, são calculados os valores da estatística $F_{parcial}$ para o novo modelo e o procedimento repete-se. O algoritmo termina quando o menor valor da estatística $F_{parcial}$ não for menor que o valor crítico da estatística.

Seleção *Forward*

Este procedimento começa com a suposição de que não há regressores no modelo a não ser a constante (β_0). O método pretende encontrar o subconjunto ótimo de regressores através da inserção de regressores no modelo um de cada vez. A primeira variável a ser introduzida é aquela que tiver maior coeficiente de correlação (em módulo) com a variável dependente y. De forma sequencial são introduzidas as variáveis com maior coeficiente de correlação parcial entre a variável resposta e a variável que se pretende incluir tendo em conta as variáveis já introduzidas.

Em cada passo é avaliado o valor da estatística $F_{parcial}$ correspondente ao novo parâmetro de regressão introduzido. Se $F_{parcial}$ for menor a um determinado valor crítico já estabelecido a variável que se acabou de introduzir é eliminada e considera-se uma nova variável até não existirem mais variáveis para introduzir no modelo.

Seleção *Stepwise*

A seleção *Stepwise* é uma modificação da seleção *Forward*, em que a cada passo todas as variáveis explicativas introduzidas no modelo são também avaliadas através do valor da estatística $F_{parcial}$. Isto é, as variáveis são introduzidas uma a uma mas em cada passo é feita uma análise das variáveis já introduzidas até aí, por forma a garantir a significância das variáveis.

3.1.8 Multicolinearidade

O uso e a interpretação de um modelo de regressão múltipla, muitas vezes depende explicita ou implicitamente nas estimativas dos coeficientes de regressão. Em algumas situações, os regressores são quase linearmente relacionados, e, nesses casos, as inferências baseadas no modelo podem ser erróneas. Quando existem dependências quase lineares entre os regressores, o problema de multicolinearidade existe. A multicolinearidade pode ter efeitos consideráveis não apenas sobre a habilidade preditiva do modelo de regressão, mas também sobre a estimação dos coeficientes de regressão e os seus testes de significância.

Existem quatro principais fontes de multicolinearidade:

1. O método de seleção dos dados;
2. Restrições sobre o modelo ou sobre a população;
3. Especificações do modelo;
4. Modelo com regressores redundantes.

Os regressores são as colunas da matriz X. Como tal uma dependência linear resultaria em uma matriz X^TX singular. Por isso, uma técnica de deteção de multicolinearidade é analisar a matriz de correlações. Examinar as correlações simples entre as variáveis explicativas é útil na deteção de dependência quase linear apenas entre pares de regressores. Quando mais do que dois regressores estão envolvidos numa dependência quase linear não existe garantia de que qualquer correlação entre pares seja grande. Por isto é importante abordar outros métodos de deteção da presença de multicolinearidade.

Os elementos diagonais da matriz $C = (X^TX)^{-1}$ são frequentemente chamados de fatores de inflação da variância (VIFs), e é um diagnóstico muito importante para a deteção de

multicolinearidade. Pode-se então definir os fatores de inflação da variância da seguinte forma:

$$\text{VIF}_j = C_{jj} = \frac{1}{1 - {R_j}^2}$$

onde ${R_j}^2$ é o coeficiente de determinação (R^2) da regressão de x_j sobre as outras variáveis explicativas. Se x_j não for muito correlacionado com os restantes regressores, o valor de $1 - {R_j}^2$ será próximo da unidade, pelo que um ou mais valores elevados de VIFs indicam multicolinearidade. A experiência diz que, se qualquer um dos VIFs for superior a 5 ou 10 é indicativo de que a estimação dos coeficientes de regressão associados a esses valores é pobre devido à multicolinearidade.

Existem várias técnicas para lidar com os problemas causados pela multicolinearidade. As abordagens gerais incluem a recolha de dados adicionais, re-especificação do modelo de regressão, bem como a utilização de métodos de estimação que não o método dos mínimos quadrados.

3.1.9 Variáveis Mudas

Variáveis Mudas ou variáveis *dummy* são variáveis que tomam somente dois valores - 0 e 1. Normalmente, o valor 1 representa a presença de algum atributo e o valor 0 a ausência do mesmo. Estas variáveis têm uma vasta gama de aplicações, algumas que irão ser discutidas neste capítulo.

Variáveis mudas que assumam somente dois valores podem ser designadas por variáveis dicotómicas. Variáveis que tenham um número finito de valores - mas mais do que dois - podem ser denominadas variáveis policotómicas.

Geralmente as variáveis policotómicas são qualitativas, mas por vezes são úteis no estudo das variáveis numéricas. Muitas vezes variáveis ordinais são tratadas como variáveis policotómicas embora não lhes seja atribuída nenhuma ordem para os níveis dessas variáveis. Na análise de regressão, por vezes é necessário de incorporar variáveis quantitativas (ou fatores). É a partir das variáveis mudas que este tipo de variáveis são introduzidas no modelo. Por exemplo, caso se pretenda modelar uma variável resposta y em função de 2 regressores, em que um deles (x_1) toma valores numéricos e o outro (x_2) é qualitativo, com x_2 a assumir dois estados A e B, considera-se a equação

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

onde $x_{i2} = 1$ se o indivíduo i pertence ao estado A, ou $x_{i2} = 0$ se o indivíduo i pertence ao estado B.

Quando existem m fatores, apenas é necessário definir $m - 1$ variáveis mudas, para identificar $m - 1$ fatores, sendo o outro identificado por complementaridade (como se pode ver no exemplo anterior).

3.1.10 Transformações

Nesta secção, apresentam-se métodos e procedimentos para a construção de modelos de regressão quando alguns dos pressupostos são violados. Não é incomum observar que quando a variável resposta e/ou os regressores são expressos na escala correta de medição, que certas violações, como a desigualdade de variância, não estão mais presentes. Idealmente, a escolha da métrica é feita pela pessoa com conhecimento de causa, mas em certas situações não existe informação disponível para tal. Nestes casos, as transformações podem ser escolhidas de forma heurística, ou por algum procedimento analítico.

Transformações de Estabilização da Variância

A suposição de variância constante é um requisito básico de análise de regressão. Uma razão comum para a violação deste pressuposto é a variável resposta y poder seguir uma distribuição de probabilidade em que a variância esteja funcionalmente relacionada com a média. Por exemplo, se y seguir uma distribuição de Poisson, pode-se considerar $y' = \sqrt{y}$ tendo em conta que a variância da raiz quadrada da Poisson é independente da média.

Quando a variável resposta é reescrita, os valores previstos encontram-se na escala transformada. Por isso, muitas vezes é necessário efetuar a conversão dos valores previstos para as unidades originais. Infelizmente, aplicando a transformação inversa diretamente nos valores previstos dá uma estimativa da mediana da distribuição da resposta em vez da média.

Transformações para Linearizar o Modelo

A hipótese de existir uma relação linear entre a variável resposta y e os regressores é habitualmente o ponto de partida na análise de regressão. Por vezes, observa-se que esta suposição é inadequada. A experiência anterior ou as considerações teóricas podem indicar que a relação entre y e os regressores não é linear. Em alguns casos, a função pode ser linearizada utilizando uma transformação adequada.

Transformações na Variável Resposta

Suponha-se que se pretende transformar y com o objetivo de corrigir a não normalidade e/ou a variância não constante dos erros do modelo. Uma classe útil de transformações é a transformação da potência y^λ, onde λ é uma parâmetro a determinar (por exemplo, se $\lambda = \frac{1}{2}$ significa que a variável resposta passa a ser $\sqrt{y}$). Box and Cox [1964] mostraram que os parâmetros do modelo de regressão e λ podem ser estimados simultaneamente através do método de máxima verosimilhança.

Transformações nas Variáveis Explicativas

Suponha-se que a relação entre y e cada um dos seus regressores não é linear mas que as suposições de normalidade, independência e variância constante são aproximadamente satisfeitas. Nesta situação, quer-se encontrar a transformação adequada sobre as variáveis explicativas para que a relação entre y e o regressor transformado seja a mais simples possível. Box and Tidwell [1962] descreveram um procedimento analítico para a determinação da transformação que deve ser executada em x.

3.2 Árvores de Regressão

Em determinadas áreas, e em muitas situações reais é frequente que as relações entre variáveis sejam fortemente não lineares e que envolvam interações muito elevadas, constituindo um impedimento à aplicação das técnicas de Análise de Regressão Linear. A construção de árvores de regressão tem fundamentalmente duas finalidades: prever a variável resposta da forma mais exata possível e/ou compreender as relações estruturais entre a variável resposta e as variáveis preditivas.

3.2.1 Introdução

Os modelos de classificação e regressão em árvores oferecem uma alternativa para estudar a relação entre uma variável dependente com uma ou mais variáveis preditivas. Existem algumas técnicas de classificação e regressão em árvore, que diferem quanto ao método usado para a divisão do conjunto de dados. As mais usadas são a CART (*Classification and Regression Trees*) e a CHAID (*Chi-Square Automatic Iterative Detection*).

A técnica CART divide os dados repetidamente e sequencialmente, de forma a que os subgrupos resultantes de cada divisão apresentem entre si a maior heterogeneidade possível e a maior homogeneidade interna. Contrariamente à técnica CART, que apenas permite partições binárias do conjunto de dados, a técnica CHAID, baseada em testes do Qui-Quadrado que são aplicados sequencialmente, permite a sua divisão em dois ou mais grupos.

Neste relatório irá ser utilizada a técnica CART, pelo que a árvore é construída repetindo sucessivas divisões dos dados, através de uma regra simples baseada numa única variável preditiva. Em cada divisão os dados são separados em dois grupos mutuamente exclusivos, o mais homogéneos, possíveis. O procedimento de divisão é repetido e aplicado separadamente a cada um dos grupos originados.

As variáveis preditivas podem ser categóricas ou numéricas e dependendo da variável dependente ser categórica ou numérica, têm-se, respetivamente Árvores de Classificação ou Árvores de Regressão. Iremos considerar y como a variável resposta e x_i com $i = 1, \cdots, p$ as variáveis preditivas onde p corresponde ao número de preditores.

O objetivo é constituir partições da variável resposta em grupos homogéneos mas mantendo a árvore razoavelmente pequena. A divisão é feita repetidamente até que se obtenha uma árvore de grande dimensão (isto é, com muitos grupos finais/terminais), que depois é

podada até se obter uma dimensão desejada. Cada grupo é caracterizado pela distribuição da variável resposta se ela for categórica ou pelo correspondente valor médio se ela for numérica, e ainda pelo tamanho e valores das variáveis preditivas que a definem.

A forma como as variáveis preditivas são usadas para formar grupos, depende do seu tipo. No caso de variáveis preditivas categóricas com 2 níveis, apenas se pode fazer uma divisão. Se a variável tiver $k(> 2)$ níveis existem $2^{k-1} - 1$ possibilidades de fazer divisões. Se as variáveis preditivas forem numéricas, a divisão baseia-se na verificação de uma condição do tipo $x_i > v$ ou $x_i \leq v$; para este valor v consideram-se todos os valores intermédios da correspondente variável preditiva. De todas as possíveis divisões, considerando todas as variáveis preditivas, seleciona-se a que torna máxima a homogeneidade dos grupos resultantes. A homogeneidade pode ser definida e avaliada de várias formas, e a sua escolha depende do tipo de variável resposta.

Uma árvore de decisão é composta por um nó raiz, um conjunto de nós interiores, e nós terminais também denominados nós folha. O nó raiz e os nós interiores, referidos coletivamente como nós não terminais, estão ligados em fases de decisão. Os nós terminais representam as classes finais. Assim, a árvore de regressão representa um conjunto de restrições ou condições que são hierarquicamente organizadas, e que são aplicadas sucessivamente a partir de uma raiz a um nó terminal ou folha da árvore.

Uma vez que, no problema em estudo, a variável dependente QUANT é numérica, iremos desenvolver as noções gerais com especial enfoque nas Árvores de Regressão.

3.2.2 Noções associadas a Árvores de Regressão

Para a construção de uma árvore de regressão, segundo Torgo (1998) é necessário ter em conta três pontos:

1. seleção do ”melhor”critério para a divisão;
2. critério de paragem;
3. atribuição de um valor aos nós folha.

Seleção do ”melhor”critério de divisão

A homogeneidade dos nós é definida por uma medida de impureza que toma o valor zero para nós completamente homogéneos e aumenta à medida que diminui a homogeneidade.

As medidas de impureza propostas na literatura (Breiman et al, 1984, Ripley, 1996, Clark e Pregisbon, 1992, entre outros), diferem entre si e dependem de se tratar de uma árvore de classificação ou de regressão.

Especificamente para árvores de regressão, Breiman et al (1984) sugeriu dois critérios para a seleção da melhor regra de divisão baseados nas estimativas dos erros obtidos, baseado no modelo de regressão: o Erro Quadrático Médio (EQM) e o Erro Absoluto Médio (EAM) também designado como Desvio Absoluto Médio. Supondo n a dimensão dos dados utilizados estas medidas definem-se da seguinte forma:

i. $\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, com $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;

ii. $\text{EAM} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - m_e \mid$, com m_e : mediana das n observações y_i.

O Erro Quadrático Médio e o Erro Absoluto Médio são medidas de impureza pois tomam o valor zero para os nós completamente homogéneos e aumentam à medida que diminui a homogeneidade. Assim, maximizar a homogeneidade é equivalente a minimizar o EQM ou EAM.

Ainda segundo Breiman et al, (1984), este critério conduz a árvores robustas mas pode ser pouco eficiente quando os dados tiverem muitos valores nulos e as variáveis explicativas forem categóricas.

Assim, procura-se o teste/regra, experimentando todas as variáveis preditivas e tomando todos os valores intermédios para v correspondentes a essas variáveis que minimizam o EQM ou EAM.

Critério de Paragem

Comece-se por definir o erro associado ao teste t, usando n elementos:

$$\text{Erro}(n_t) = \frac{1}{n}\sum_{i \in P_n}(y_i - \bar{y}_n)^2.$$

O critério de paragem pode assentar num dos seguintes procedimentos:

→ Continua-se a fazer a divisão até que o "ganho"devido à divisão adicional seja inferior a um valor pré-definido. Este procedimento tem algumas fragilidades: se a regra de paragem é baseada num número muito pequeno conduz a árvores excessivamente grandes - problema de "*overfitting*"; se o número for muito elevado, corre-se o risco de a árvore não expressar devidamente as interações entre as variáveis explicativas.

→ Para-se o crescimento de o erro existente antes da divisão for inferior ao erro associado correspondente após a divisão. Nesta metodologia, analisando a figura 3.2 temos que o erro antes da divisão é definido por $\text{Erro}_{\text{antes}} = \text{Erro}(n_t)$ e o erro após a divisão por $\text{Erro}_{\text{após}} = \frac{n_{t,e}}{n_t}\text{Erro}(n_{t,e}) + \frac{n_{t,d}}{n_t}\text{Erro}(n_{t,d})$. Em cada divisão comparam-se os erros correspondentes e se $\text{Erro}_{\text{após}} \geq \text{Erro}_{\text{antes}}$, para-se o crescimento da árvore.

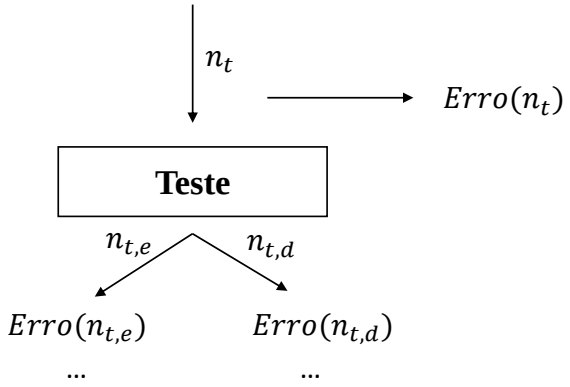

Figura 3.2: Melhor teste para um nó.

É relevante ter em conta que com o processo de subdivisão sucessiva da amostra temos cada vez menos dados e isso levanta o problema da falta de fiabilidade das estimativas obtidas para os erros e o problema de *overfitting*. De forma a resolver este problema podem-se aplicar as seguintes estratégias:

i. Usar melhores estimativas para o erro;

ii. Incluir, no critério de paragem do crescimento das árvores, métodos de avaliação mais fiáveis;

iii. Em vez de parar o crescimento da árvore, constrói-se uma árvore de grande dimensão e depois poda-se.

3.2.3 Melhorar as Estimativas do Erro

De forma a melhorar as estimativas do erro têm-se os seguintes métodos:

- **Método de Re-amostragem** (o método de "Holdout")

  É o método mais usual e consiste em dividir a amostra em dois subconjuntos mutuamente exclusivos, um para estimação, isto é, para obter as árvores, e outro para

obter as estimativas do erro. Habitualmente, uma das amostras contém $\frac{2}{3}$ das observações que é utilizada para obter a árvore; e a outra amostra com $\frac{1}{3}$ das observações é usada para obter as estimativas do erro. O método tem a vantagem de usar amostras independentes para obter a árvore e as estimativas mas também pode ser uma desvantagem pois fica-se com menos dados para a obtenção da árvore de regressão.

Esta abordagem é indicada quando está disponível uma grande quantidade de dados. Caso a dimensão da amostra seja pequena, o erro calculado na predição pode sofrer uma grande variação.

- **Método da Validação Cruzada**

  O método divide os dados em k subconjuntos com a mesma dimensão. De cada vez, usam-se $k-1$ conjuntos para estimação e um para validação, e repete-se o procedimento k vezes. No final, a estimativa obtida para os erros é a média das k estimativas parciais. Usualmente considera-se $k = 10$.

  Este método dá-nos estimativas muitos "fiáveis"mas pode ser computacionalmente "pesado", mas é o mais utilizado.

- **Método *leave-one-out***

  Caso particular da validação cruzada onde o número de subconjuntos coincide com o número de elementos da amostra. Este método é vantajoso quando temos poucos dados disponíveis.

3.2.4 Critérios de Paragem mais eficazes

Para dispormos de critérios de paragem mais eficazes deve-se melhorar as estimativas do erro usadas. Por exemplo, no cálculo do erro (medido pelo EQM) pode-se usar o método de *Holdout*. O mesmo procedimento pode ser efetuado para calcular o erro após a divisão, $\text{Erro}_{\text{após}}$, isto é, substituir os erros anteriormente especificados pelos respetivos erros obtidos através do método de *Holdout* ou do método da validação cruzada.

Falta-nos responder a uma questão importante: "Como e quando parar?". Deve-se parar o crescimento da árvore se $\text{Erro}_{\text{após}} \geq \text{Erro}_{\text{antes}}$, isto significa que fazer a divisão do nó já não seria proveitoso quanto ao erro.

3.2.5 Poda da Árvore

Segundo Breiman (1984), para prevenir o problema do *overfitting* e melhorar as previsões deve-se realizar a poda da árvore. A poda das árvores de regressão é um procedimento "*standard*"neste tipo de metodologias cujo objetivo principal é o de proporcionar um melhor compromisso entre a simplicidade e compreensibilidade das árvores e a sua capacidade preditiva. Ao fazer-se a poda da árvore estamos a obter resultados melhores em termos do erro. Dá-nos também a possibilidade de "inspecionar"um conjunto de modelos alternativos com diferente compromisso do tamanho/erro.

Existem duas formas de se proceder à poda da árvore:

→ Realizando uma pré-poda, isto é, utilizar critérios de paragem mais eficientes. Neste contexto, fazer a poda da árvore, consiste em

- obter uma árvore exageradamente grande;
- gerar uma sequência de sub-árvores;
- escolher a "melhor"sub-árvore, usando métodos de avaliação "fiáveis"para escolher o modelo final.

→ Realizando uma pós-poda, isto é, cortando ramos desnecessários depois de estar completa.

O método de poda é utilizado na maioria dos sistemas/*packages* informáticos desenvolvidos (como por exemplo, a *package tree* e *rpart*). No entanto, em problemas de grande dimensão é pouco eficiente computacionalmente.

A secções 3.2.4 e 3.2.5 encontram-se interligadas entre si. Se utilizarmos critérios de paragem mais eficazes e procedermos à poda da árvore estamos a realizar uma paragem mais eficaz. Isto origina resultados comparáveis em termos de erro e uma grande eficiência em problemas de grande dimensão. No entanto, o crescimento da árvore pode parar "cedo de mais"e deixa-se de obter uma sequência de modelos alternativos.

É importante observar que o procedimento deduzido baseia-se no EQM, mas também poderia ser avaliado em termos do EAM, que seria mais adequado quando existissem distâncias enviesadas e na presença de *outliers*. No entanto, é mais difícil em termos de otimização computacional, pelo que, apesar de mais robusto, poderá ser menos eficiente.

Após a definição das noções gerais sobre árvores de regressão, iremos abordar duas *packages* do *software* estatístico R, *rpart* e *tree*, que podem ser usadas para a construção de árvores de regressão baseadas no método CART. Estas *packages* são as mais utilizadas para a construção de árvores de regressão. No entanto, é necessário ter cuidado pois podem produzir resultados muito diferentes.

Seguidamente, iremos distinguir a *package tree* da *package rpart*.

3.2.6 *Package tree*

A *package tree* é a primeira *package* do R para árvores de classificação e regressão. A função que constrói a árvore pode incluir um critério de paragem pré-definido. No entanto, no caso de não ser incluído pára-se a divisão se a quantidade adicional da *deviance* não for significativa.

A função que efetua a poda da árvore utiliza a validação cruzada como método.

3.2.7 *Package rpart*

A *package rpart* fundamenta-se num algoritmo mais robusto. O algoritmo *rpart* do R baseia-se numa medida do erro da complexidade da árvore. Esta medida é definida da seguinte forma:

$$\mathrm{EC}_\alpha(T) = \mathrm{Erro}(T) + \alpha \cdot \#\tilde{T},$$

onde $\mathrm{Erro}(T)$ designa a estimativa do erro, α corresponde ao parâmetro de complexidade que define o custo de cada folha e $\#\tilde{T}$ é o número de folhas da árvore.

Em vez de se controlar o número de divisões diretamente, este algoritmo controla indiretamente, através de uma quantidade α (parâmetro de complexidade - designado no pacote como cp) que coloca um custo adicional a cada divisão. O aumento do "custo"à medida que a árvore se torna mais complexa é analisado conjuntamente com o critério de paragem. Um grande valor para α leva a uma árvore pequena, enquanto que um valor pequeno gera uma árvore complexa.

Assim, o principal objetivo do algoritmo é minimizar a complexidade da árvore. Para isso, a divisão é feita procurando o valor máximo do parâmetro α que torna

$$\mathrm{EC}_\alpha(T) < xerror + xstd, \tag{3.10}$$

onde $xerror$ é o valor mínimo do erro da validação cruzada e $xstd$ o correspondente desvio padrão.

Em síntese, para a *package rpart* o valor do parâmetro de complexidade deve ser suficientemente pequeno de forma a que o erro da validação cruzada atinja o seu mínimo. Tendo sido identificada a árvore ideal (isto é, com o erro de validação cruzada mínimo), o número de divisões pode ainda ser podado.

Capítulo 4

Aplicação ao Caso de Estudo

Esta secção tem como objetivo a apresentação do problema proposto pelo Porto de Aveiro, bem como o desenvolvimento de uma abordagem metodológica para o problema em questão e a análise dos resultados relevantes.

4.1 Definição do Problema

Durante o período de estágio na administração portuária foi proposto pelo Dr. Luís Sousa um estudo estatístico incidente na exportação do cimento. Segundo fontes do Porto de Aveiro, a exportação de cimento tem vindo a aumentar nos últimos anos e é relevante para o porto entender quais a variáveis que justificam esse aumento.

Em suma, pretende-se com este trabalho:

- identificar as variáveis que podem influenciar o aumento da quantidade exportada de cimento no Porto de Aveiro;
- aferir se as variáveis justificam o aumento da quantidade exportada de cimento no Porto de Aveiro.

4.2 Metodologia e Dados

A seleção da metodologia de análise de um conjunto de dados é de vital importância para a concretização dos objetivos definidos, no contexto de um problema específico. O número de variáveis envolvidas, as respetivas escalas de medida e as relações de dependência ou interdependência entre elas são aspetos relevantes para essa decisão.

Através dos registos do porto verifica-se que a exportação do cimento tem a sua origem em 2005. Consequentemente, o período analisado foi de 2005 até ao segundo trimestre de 2014, onde os dados se encontram agrupados por trimestre.

O Porto de Aveiro nos últimos anos tem beneficiado com o aumento da exportação de cimento. Tendo isto como motivação é relevante identificar as variáveis que mais influenciam a quantidade de cimento exportada. Para isso, segui a orientação do Dr. Luís Sousa e de algumas entidades que trabalham na administração do porto. As principais respostas obtidas sobre as variáveis que influenciam a exportação do cimento foram: as dimensões dos navios, as acessibilidades marítima, terrestre e ferroviária e a produtividade do porto.

Relativamente à dimensão dos navios foi apurado que é importante considerar as seguintes três variáveis: a arqueação bruta (tonelagem bruta) do navio que corresponde a todos os volumes interiores fechados do navio, o comprimento do navio e o calado do navio que é a profundidade em que cada navio está submerso na água. Como a distribuição dos dados correspondentes é assimétrica é utilizada como unidade de medida central a mediana amostral. Definem-se então as medidas trimestrais: GTm como a mediana da arqueação bruta do navio, LOAm como a mediana do comprimento do navio e CALm como a mediana do calado do navio.

Quanto à acessibilidade marítima chegou-se à conclusão que é importante considerar a variável cota de serviço máxima do canal principal de navegação, denominada COTA, que nos indica a profundidade máxima que o canal permite. Esta variável pode tomar os valores: $\{-8; -10,5; -12,5\}$. No entanto, o que é relevante não é se a cota de serviço máxima toma cada um dos valores mas sim em qual dos níveis se encontra.

No que concerne à acessibilidade ferroviária será considerada uma variável muda que nos indica se na altura existia ou não a ferrovia em funcionamento. Esta variável pode definir-se

da seguinte forma:

$$\text{AF} = \begin{cases} 1, & \text{se a ferrovia existir} \\ 0, & \text{caso contrário} \end{cases}.$$

A acessibilidade terrestre não foi representada por nenhuma variável pois não existiu nenhuma alteração significativa no período considerado dos dados.

No que respeita à produtividade do porto considerámos a variável produtividade média de atendimento ([2]):

$$\text{PROD} = \frac{\text{quantidade de mercadoria movimentada pelo navio (em KG)}}{\text{número de horas atracado (em horas)}},$$

que determina se o porto necessita ou não de construção de mais infra-estruturas para satisfazer a procura.

Uma variável que deveria ser também considerada é o nível de serviço ([2]) que estabelece a relação entre os tempos de espera e os tempos de atendimento em percentagem (%). A variável é definida da seguinte forma:

$$\text{Nível de serviço} = \frac{\text{tempo de espera do navio}}{\text{tempo de atendimento do navio}}.$$

No entanto, esta variável não foi considerada por falta de dados.

Um dos objetivos do trabalho é estudar a relação que a quantidade de cimento exportada tem com as variáveis consideradas. Assim sendo, faz sentido utilizar a Regressão Múltipla. A análise de regressão múltipla é uma técnica estatística que pode ser usada para analisar a relação entre uma única variável dependente e múltiplas variáveis independentes (preditoras). O *software* estatístico utilizado para a análise do modelo de regressão linear múltipla foi o SPSS (*Statistical Package for Social Sciences*). Por outro lado, como a análise deste problema através de modelos de regressão linear múltipla incluiu transformações da variável dependente, e atendendo à possibilidade de interações elevadas entre as variáveis, utiliza-se também a metodologia das Árvores de Regressão. Neste caso utilizou-se o pacote *tree*, *rpart* e *rpart.plot* do R.

4.3 Caso de Estudo: Modelo de Regressão Linear Múltipla

Define-se como variável resposta/dependente a quantidade de cimento exportada em KG por trimestre, denominada QUANT. As variáveis independentes/explicativas serão: GTm, LOAm, CALm, COTA, AF, PROD.

Para incluir a variável categórica COTA no modelo de regressão, e tomando como referência o valor de profundidade $-12,5$, definem-se apenas duas variáveis mudas C1 e C2 da seguinte forma:

$$\text{C1} = \begin{cases} 1, & \text{se a COTA=-8} \\ 0, & \text{caso contrário} \end{cases},$$

$$\text{C2} = \begin{cases} 1, & \text{se a COTA=-10,5} \\ 0, & \text{caso contrário} \end{cases}.$$

É óbvio que se C1 $= 0$ e C2 $= 0$ então a variável COTA $= -12,5$. Assim, as variáveis explicativas passam a ser: GTm, LOAm, CALm, C1, C2, AF, PROD.

Após a definição da variável resposta e das variáveis explicativas, pode-se fazer uma análise preliminar dos gráficos de dispersão da variável dependente $y =$ QUANT versus cada uma das variáveis explicativas (contínuas) como se pode observar na figura 4.1.

Desta análise pode-se notar que os vários regressores em causa, o GTm, LOAm, CALm e PROD não parecem estar muito relacionadas com a variável QUANT. No entanto, verifica-se que algumas das correlações entre regressores são elevadas como a tabela 4.1 nos mostra. Pelo que não é de estranhar que os gráficos de dispersão entre y e cada regressor não fossem muito informativos.

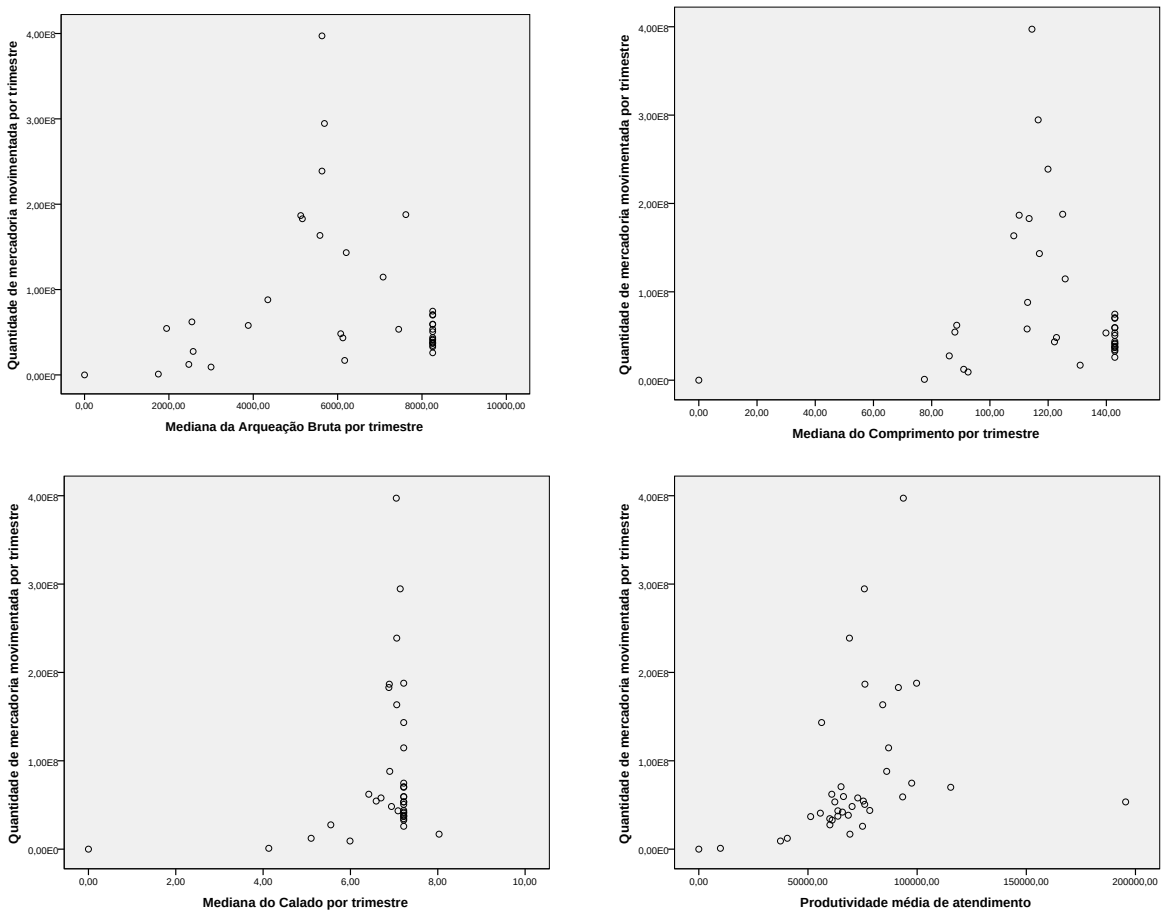

Figura 4.1: Gráficos de dispersão da variável dependente versus cada um dos regressores.

Correlations									
		QUANT	GTm	LOAm	CALm	AF	PROD	C1	C2
Pearson Correlation	QUANT	1,000	0,025	0,025	0,245	0,620	0,302	-0,508	-0,117
	GTm	0,025	1,000	0,930	0,724	-0,119	0,476	-0,117	0,165
	LOAm	0,025	0,930	1,000	0,888	-0,065	0,513	-0,141	0,192
	CALm	0,245	0,724	0,888	1,000	0,193	0,563	-0,294	0,219
	AF	0,620	-0,119	-0,065	0,193	1,000	0,221	-0,767	0,412
	PROD	0,302	0,476	0,513	0,563	0,221	1,000	-0,525	0,391
	C1	-0,508	-0,117	-0,141	-0,294	-0,767	-0,525	1,000	-0,727
	C2	-0,117	0,165	0,192	0,219	0,412	0,391	-0,727	1,000

Tabela 4.1: Matriz de correlações do modelo.

Numa primeira instância, considerou-se o modelo de regressão linear múltipla para a variável dependente QUANT e considerando as variáveis independentes GTm, LOAm, CALm,

C1, C2, AF e PROD. Aquando desta análise notaram-se diversos problemas, como que se pode comprovar através de alguns dos resultados obtidos apresentados no apêndice C.

A tabela C.1 apresenta os resultados da aplicação do modelo de regressão referido em função do R^2.

Um primeiro indício foi as variáveis AF e PROD não serem significativas para o modelo, como se depreende da tabela C.2, tendo em conta que o aparecimento da ferrovia foi uma das causas do aumento da exportação de cimento não faz sentido o resultado contraditório. Ainda da análise da tabela referida, notou-se também multicolinearidade entre as variáveis GTm, LOAm e CALm, o que nos indica que as variáveis possuem uma relação linear entre si. Este problema origina uma dificuldade na interpretação dos valores dos coeficientes estimados.

Pelo *PP-plot* (Normal) dos resíduos (figura C.1) e pelos testes formais Kolmogorov-Smirnov e Shapiro-Wilk (tabela C.3) verificou-se que o pressuposto da normalidade dos resíduos falha, o que nos indica que a regressão não está a ser adequada.

Através de uma análise ao gráfico de resíduos versus os valores preditos, ilustrado na figura C.2, nota-se uma tendência e um crescimento nos resíduos, o que nos levanta a suspeita de existência de heterocedasticidade. De forma a verificar a existência de heterocedasticidade foram utilizados os testes de Breusch-Pagan e Koenker. Através de uma MACRO[1] já existente (ver apêndice D), verificou-se que os seus p-values (aproximadamente 0 e 0,0444, respetivamente) são inferiores aos níveis de significância $\alpha = 0.05$ e $\alpha = 0.1$. Assim sendo, rejeita-se a hipótese nula de homocedasticidade, por isso, temos motivos para concluir que existe heterocedasticidade.

Pelo facto de existirem tantos problemas optou-se por nos focarmos na questão da falta de homocedasticidade. Quando existe heterocedasticidade um dos motivos que pode estar na sua origem é o modelo que está a ser usado não ser adequado ao problema, e neste caso particular, não haver uma dependência linear entre a variável resposta e as variáveis explicativas. Este problema pode ser solucionado através de uma transformação adequada na variável resposta. Fez-se então a seguinte transformação:

$$\text{raizQUANT} = \sqrt{\text{QUANT}}.$$

O modelo linear passa agora a ter a variável raizQUANT como variável resposta e as

[1] **Fonte:** http://www.spsstools.net/Syntax/RegressionRepeatedMeasure/Breusch-PaganAndKoenkerTest.txt

variáveis explicativas: GTm, LOAm, CALm, AF, C1, C2 e PROD.

As estatísticas descritivas das variáveis em causa encontram-se na tabela 4.2 (excluindo as variáveis mudas).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
raizQUANT	38	,00	19930,51	8123,3608	4181,51555
Mediana da Arqueação Bruta por trimestre	38	,00	8254,00	6160,6053	2388,83699
Mediana do Comprimento por trimestre	38	,00	142,91	121,1147	28,63631
Mediana do Calado por trimestre	38	,00	8,03	6,7347	1,31516
Produtividade média de atendimento	38	,00	195499,33	71840,2355	30144,98935
Valid N (listwise)	38				

Tabela 4.2: Estatísticas Descritivas.

A matriz de correlações permite-nos verificar se existe uma estrutura de correlação entre as variáveis. Tal como se procedeu anteriormente, é útil examinar a matriz de correlações antes de aplicar a regressão linear múltipla. Através da tabela 4.3, nota-se que a variável raizQUANT está bastante correlacionado com as variáveis C1 e AF. As variáveis GTm, LOAm e CALm estão fortemente correlacionadas entre si, o que indicia possível existência de multicolinearidade entre elas. A variável C1 e C2 encontram-se correlacionadas entre si.

Correlations									
		raizQUANT	GTm	LOAm	CALm	C1	C2	AF	PROD
Pearson Correlation	raizQUANT	1,000	0,182	0,211	0,448	-0,592	0,015	0,680	0,444
	GTm	0,182	1,000	0,930	0,724	-0,117	0,165	-0,119	0,476
	LOAm	0,211	0,930	1,000	0,888	-0,141	0,192	-0,065	0,513
	CALm	0,448	0,724	0,888	1,000	-0,294	0,219	0,193	0,563
	C1	-0,592	-0,117	-0,141	-0,294	1,000	-0,727	-0,767	-0,525
	C2	0,015	0,165	0,192	0,219	-0,727	1,000	0,412	0,391
	AF	0,680	-0,119	-0,065	0,193	-0,767	0,412	1,000	0,221
	PROD	0,444	0,476	0,513	0,563	-0,525	0,391	0,221	1,000

Tabela 4.3: Matriz de correlações do modelo.

Utilizando o *software* estatístico SPSS, ao efetuar a análise de regressão com todas as variáveis obtêm-se as seguintes tabelas de resultados:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,931[a]	,867	,836	1695,33153	1,484

a. Predictors: (Constant), Produtividade média de atendimento, Ferrovia, c2, Mediana da Arqueação Bruta por trimestre, Mediana do Calado por trimestre, c1, Mediana do Comprimento por trimestre

b. Dependent Variable: raizQUANT

Tabela 4.4: Resumo do modelo.

Os valores do coeficiente de determinação (ver tabela 4.4) ajustado $(0,836)$ ou não ajustado $(0,867)$ permitem considerar como válido o modelo de regressão linear múltipla para explicar possíveis variações na quantidade de cimento exportada, isto é, aproximadamente 84% da variabilidade existente no modelo é explicada pela regressão múltipla aplicada. Comparando a estimativa do desvio padrão da variável raizQUANT $(4181,52)$ com a estimativa do desvio padrão dos erros $(1695,33$ como é possível observar na tabela 4.4), pode-se observar uma diminuição significativa da variabilidade. Assim, o modelo de regressão parece contabilizar uma parte significativa da variabilidade das observações.

Com o objetivo de detetar a presença de autocorrelação (dependência) nos resíduos utilizou-se o teste de Durbin-Watson. O valor da estatística de teste (ver tabela 4.4) é $d = 1,484$. Como $d_L = 0,913 < d < d_U = 1,735$, o teste é inconclusivo, o que significa que não é possível concluir nada quanto à autocorrelação dos resíduos para este caso.

Analisando a tabela ANOVA (ver tabela 4.5) temos um *p-value* aproximadamente nulo para o teste da significância da regressão. Dessa forma confirma-se que o modelo de regressão parece ser adequado. Contudo, este teste apenas nos permite concluir que algumas variáveis explicativas são de facto importantes para explicar a variabilidade da variável resposta.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5,607E8	7	80103314,988	27,870	,000[a]
	Residual	86224469,477	30	2874148,983		
	Total	6,469E8	37			

a. Predictors: (Constant), Produtividade média de atendimento, Ferrovia, c2, Mediana da Arqueação Bruta por trimestre, Mediana do Calado por trimestre, c1, Mediana do Comprimento por trimestre

b. Dependent Variable: raizQUANT

Tabela 4.5: Teste de significância da regressão usando a ANOVA.

A partir da tabela 4.6 é possível verificar quais as variáveis explicativas que são significativas para o modelo. Verifica-se que a variável PROD se mostra não significativa para o modelo, tendo em conta que o seu *p-value*=0.252 é superior a qualquer nível de significância usual.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	6206,723	2108,690		2,943	,006	1900,203	10513,244		
	Mediana da Arqueação Bruta por trimestre	,963	,428	,550	2,250	,032	,089	1,838	,074	13,471
	Mediana do Comprimento por trimestre	-135,985	58,427	-,931	-2,327	,027	-255,309	-16,660	,028	36,038
	Mediana do Calado por trimestre	2258,636	751,444	,710	3,006	,005	723,981	3793,290	,080	12,573
	c1	-5787,887	1431,658	-,693	-4,043	,000	-8711,722	-2864,052	,151	6,606
	c2	-6012,913	894,189	-,719	-6,724	,000	-7839,090	-4186,736	,388	2,577
	Ferrovia	2382,157	1090,587	,287	2,184	,037	154,881	4609,434	,257	3,888
	Produtividade média de atendimento	,016	,014	,115	1,168	,252	-,012	,044	,460	2,172

Tabela 4.6: Estimação dos coeficientes, intervalos de confiança e teste da colinearidade.

Assim, considerou-se pertinente eliminar a variável PROD do modelo. Embora, segundo o porto de Aveiro a variável PROD seja a melhor medida para calcular a produtividade das operações de movimentação das mercadorias nos Portos, muito provavelmente os tempos de espera dos navios no porto influenciam negativamente os resultados obtidos. Sendo assim, considera-se o modelo de regressão linear múltipla com variável resposta raizQUANT e variáveis explicativas: GTm, LOAm, CALm, C1, C2 e AF.

Obtém-se os seguintes resultados:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,928[a]	,861	,834	1705,26713	1,438

a. Predictors: (Constant), Ferrovia, Mediana do Comprimento por trimestre, c2, c1, Mediana do Calado por trimestre, Mediana da Arqueação Bruta por trimestre

b. Dependent Variable: raizQUANT

Tabela 4.7: Resumo do modelo.

Os valores do coeficiente de determinação (ver tabela 4.7) ajustado $(0,834)$ ou não ajustado $(0,861)$ permitem considerar como válido o modelo de regressão linear múltipla para explicar possíveis variações na quantidade de cimento exportada, isto é, aproximadamente 83%

da variabilidade existente no modelo é explicada pela regressão múltipla aplicada. Comparando a estimativa do desvio padrão da variável raizQUANT (4181, 52) com a estimativa do desvio padrão dos erros (1705, 27 como é possível observar na tabela 4.7), pode-se observar uma diminuição significativa da variabilidade. Assim, o modelo de regressão parece contabilizar uma parte significativa da variabilidade das observações.

Analisando a tabela ANOVA (ver tabela 4.8) temos um *p-value* aproximadamente nulo para o teste da significância da regressão. Dessa forma confirma-se que o modelo de regressão parece ser adequado. Contudo, este teste apenas nos permite concluir que algumas variáveis explicativas são de facto importantes para explicar a variabilidade da variável resposta.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5,568E8	6	92800276,410	31,913	,000[a]
	Residual	90146015,936	31	2907935,998		
	Total	6,469E8	37			

a. Predictors: (Constant), Ferrovia, Mediana do Comprimento por trimestre, c2, c1, Mediana do Calado por trimestre, Mediana da Arqueação Bruta por trimestre

b. Dependent Variable: raizQUANT

Tabela 4.8: Teste de significância da regressão usando a ANOVA.

Com base na tabela 4.9 constata-se que todas as variáveis explicativas, com exceção da ferrovia, se mostram significativas para o modelo em causa, ao nível de significância 0.05. É importante notar que a variável AF já é significativa para o modelo a um nível de significância $\alpha > 0.068$, pelo que será considerada. Pela análise dos coeficientes estandardizados, verifica-se que, de entre as variáveis explicativas numéricas a mediana do comprimento é a que apresenta maior peso preditivo, enquanto que a mediana da arqueação bruta é a que tem menor peso. Assim sendo, pode-se escrever a equação de regressão estimada:

$$\begin{aligned}\text{raizQUANT} =& 6925,236 + 1,023\text{GTm} - 142,437\text{LOAm} + 2465,09\text{CALm} \\ & - 6547,529\text{C1} - 6109,977\text{C2} + 1958,979\text{AF}.\end{aligned}$$

A equação anterior estima a raiz quadrada da quantidade de cimento exportada (em kilogramas) através das variáveis explicativas.
Na tabela 4.9 podem-se ainda observar intervalos de confiança a 95% para as estimativas

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	6925,236	2028,799		3,413	,002	2787,474	11062,999		
	Mediana da Arqueação Bruta por trimestre	1,023	,428	,584	2,391	,023	,150	1,895	,075	13,281
	Mediana do Comprimento por trimestre	-142,437	58,506	-,975	-2,435	,021	-261,762	-23,112	,028	35,716
	Mediana do Calado por trimestre	2465,090	734,643	,775	3,355	,002	966,776	3963,404	,084	11,877
	c1	-6547,529	1282,901	-,783	-5,104	,000	-9164,023	-3931,035	,191	5,243
	c2	-6109,977	895,537	-,731	-6,823	,000	-7936,436	-4283,517	,391	2,555
	Ferrovia	1958,979	1034,683	,236	1,893	,068	-151,271	4069,230	,289	3,459

a. Dependent Variable: raizQUANT

Tabela 4.9: Estimação dos coeficientes, intervalos de confiança e teste da colinearidade.

dos coeficientes de regressão, dados por:

$$\text{Mediana da Arqueação Bruta (GTm)} : (0,15; 1,895);$$
$$\text{Mediana da Comprimento (LOAm)} : (-261,762; -23,112);$$
$$\text{Mediana do Calado (CALm)} : (966,776; 3963,404);$$
$$\text{C1} : (-9164,023; -3931,035);$$
$$\text{C2} : (-7936,436; -4283,517);$$
$$\text{Ferrovia (AF)} : (-151,271; 4069,230).$$

Através dos valores VIF apresentados na tabela 4.9 conclui-se que as variáveis GTm, LOAm e CALm são multicolineares. No entanto, se fizermos a regressão linear múltipla utilizando o método de seleção *stepwise* das variáveis verifica-se que GTm e LOAm são eliminadas do modelo resolvendo assim o problema de multicolinearidade, conforme se pode ver na tabela E.2 (Apêndice E). É de referir que o valor do coeficiente de determinação R^2 não se altera muito quando as variáveis GTm e LOAm são eliminadas do modelo (ver tabela E.1). Contudo, por informações recolhidas no Porto de Aveiro, optou-se por incluir as variáveis GTm e LOAm por serem consideradas relevantes no contexto do problema em análise.

Por forma a validar os pressupostos da regressão construíu-se um histograma dos resíduos, um PP-plot (Normal) dos resíduos, os testes de Kolmogorov-Smirnov e Shapiro-Wilk e um gráfico de dispersão dos resíduos estandardizados contra os valores preditos.
O histograma (figura 4.2) e o PP-plot (figura 4.3) indicam-nos que não temos motivos para rejeitar a distribuição Normal. De modo a verificar esta conclusão fez-se os testes

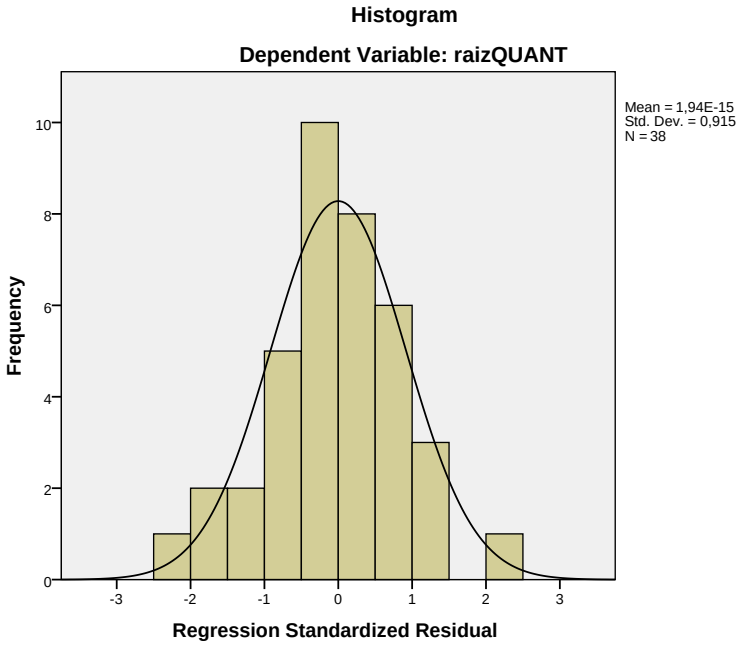

Figura 4.2: Histograma.

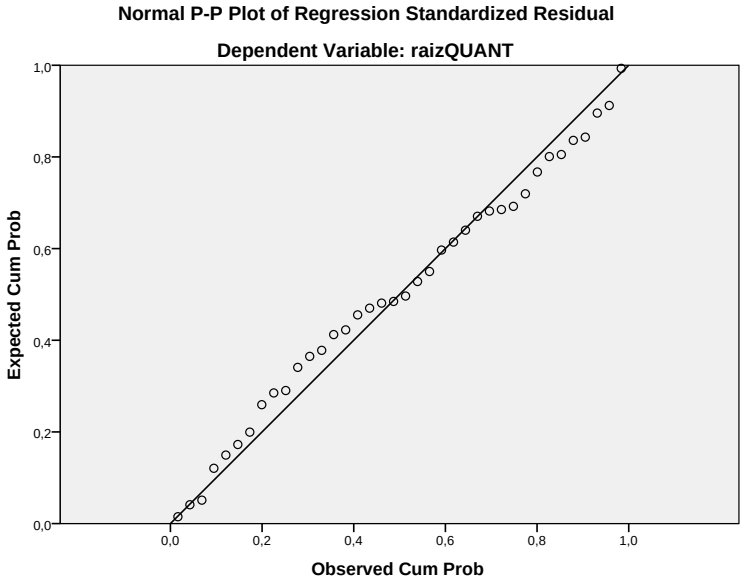

Figura 4.3: *PP-plot* (Normal) dos resíduos.

de ajustamento de Kolmogorov-Smirnov e Shapiro-Wilk (tabela 4.10). Os *p-values* dos testes (0,2 e 0,885, respetivamente) apontam para a não rejeição da hipótese nula, isto é, a normalidade dos resíduos não é rejeitada.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,064	38	,200*	,985	38	,885

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Tabela 4.10: Resultado dos testes de Kolmogorov-Smirnov e de Shapiro-Wilk.

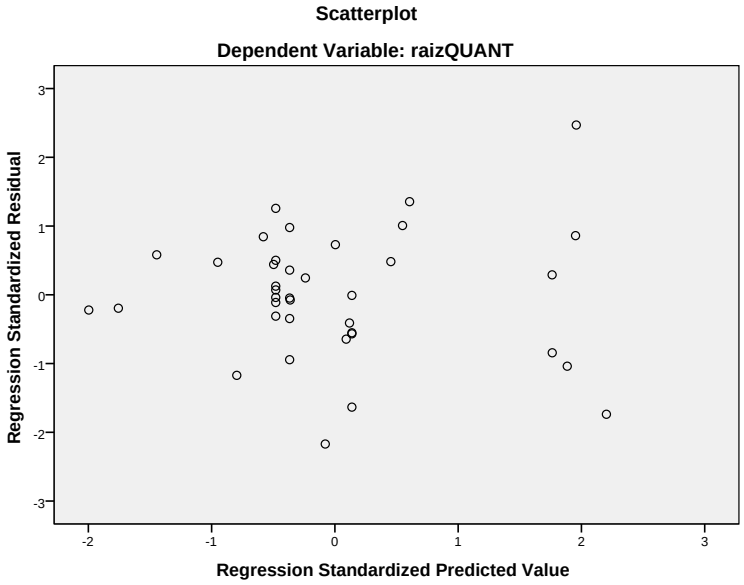

Figura 4.4: Gráfico de resíduos versus valores preditos.

O gráfico dos resíduos contra os valores preditos (figura 4.4) parece não por em causa a independência e igualdade de variâncias dos resíduos, uma vez que os valores não evidenciam uma tendência óbvia. No entanto, é apropriado verificar essas hipóteses através de testes formais. Para testar a homocedasticidade e realizando os testes de Breusch-Pagan e de Koenker obtém-se os *p-values* 0,0469 e 0,1249, respetivamente. Tendo em consideração que o teste de Koenker é mais robusto para dimensões pequenas pode-se concluir que para qualquer nível de significância usual a hipótese nula de homocedasticidade não é rejeitada. Por isso, não existem motivos para rejeitar a igualdade das variâncias dos resíduos.

Com o objetivo de detetar a presença de autocorrelação (dependência) nos resíduos utilizou-se o teste de Durbin-Watson. O valor da estatística de teste (ver tabela 4.4) é $d = 1,438$. Pelo facto de $0,966 < d < 1,658$, o teste é inconclusivo.

Como se viu, a relação encontrada para modelar a raizQUANT e GTm, LOAm, CALm,

C1, C2 e AF, é dada através de

$$\begin{aligned}\text{raizQUANT} =& 6925,236 + 1,023\text{GTm} - 142,437\text{LOAm} + 2465,09\text{CALm} \\ & - 6547,529\text{C1} - 6109,977\text{C2} + 1958,979\text{AF}.\end{aligned}$$

Sabemos que para se fazer uma adequada interpretação dos coeficientes estimados associados às variáveis mediana da arqueação bruta, mediana do comprimento e mediana do calado do modelo convém não esquecer que a interpretação de cada assenta no pressuposto de que as restantes variáveis explicativas, associadas a cada um dos outros parâmetros, se mantêm constantes. Assim, a cada acréscimo de uma unidade na variável CALm a resposta média na variável $\sqrt{\text{QUANT}}$ aumentará 2465,09 unidades, mantendo constantes todas as restantes variáveis.

Por outro lado, relativamente à variável mediana do comprimento, estima-se que em média o aumento de uma unidade nesta variável explicativa conduz a uma diminuição de $142,437$ unidades na variável $\sqrt{\text{QUANT}}$, mantendo constantes as outras variáveis explicativas.

As variáveis mediana da arqueação bruta, mediana do calado e existência de ferrovia estabelecem com $\sqrt{\text{QUANT}}$ uma relação direta, enquanto que as restantes, uma relação inversa.

Uma vez que a variável muda de referência relativamente à cota de serviço máxima é a que corresponde ao valor $-12,5$, é natural que a relação estabelecida entre C1 e C2 com $\sqrt{\text{QUANT}}$, reforça que a quantidade de cimento exportada aumenta com a profundidade máxima do principal canal de navegação.

4.4 Caso de Estudo: Árvores de Regressão

Nesta secção vamos usar a metodologia das Árvores de Regressão aplicada ao nosso caso de estudo. A variável resposta é a variável QUANT e as variáveis preditivas: GTm, LOAm, CALm, AF, COTA e PROD. Uma breve descrição de cada uma das variáveis utilizadas encontra-se na tabela 4.11.

Variável	Tipo	Valores
QUANT	N	0 – 397225127 KG
GTm	N	0 – 8254 KG
LOAm	N	0 – 142,91 m
CALm	N	0 – 8,03 m
AF	C	0 (se não existir ferrovia), 1 (se existir ferrovia)
COTA	N	{-8; -10,5; -12,5}
PROD	N	0 – 195499,33

Tabela 4.11: Descrição das variáveis usadas no estudo da exportação de cimento. O tipo das variáveis é designado por N = variáveis numéricas e C = variáveis categóricas.

O primeiro passo é fazer a leitura dos dados, importando o ficheiro DadosCimentoR.txt como se pode observar no apêndice F. Para a construção e análise da árvore de regressão vamos utilizar a *package rpart* e *tree*.

4.4.1 Aplicação da *package rpart*

A função *rpart* permite-nos obter a árvore. Como a nossa variável resposta é uma variável contínua é utilizado o *method="anova"*. Aplicando esta função obtém-se o seguinte *output*:

```
n= 38

node), split, n, deviance, yval
      * denotes terminal node

1) root 38 2.704478e+17  83013930
  2) AF< 0.5 21 7.863676e+15  35936060
    4) LOAm< 139.81 9 2.383056e+15  20502290 *
```

```
   5) LOAm>=139.81 12 1.728951e+15  47511380 *
 3) AF>=0.5 17 1.585472e+17 141168900 *
```

A primeira linha diz que a dimensão da amostra é 38, isto é, foram utilizadas 38 observações neste estudo (n=38). A segunda linha é uma legenda das linhas abaixo: *node* corresponde ao número do nó, *split* é a regra de divisão, n é o número de observações utilizadas para a divisão, *deviance* é a soma dos quadrados dos resíduos, SS_{Res}, e *yval* é o valor médio a adotar como resposta. Quando o algoritmo já não pode ser mais divido em nós, obtemos uma folha. O símbolo $*$ denota quando um nó corresponde a uma folha como se pode verificar na linha 3.

Assim, no caso do primeiro nó (denominado a raíz da árvore), pode-se observar que existem 38 observações. Este nó é dividido a partir da variável AF. Obtêm-se então dois subgrupos: o primeiro (nó 2, AF<0.5) é composto por 21 observações com *deviance* igual a 7.863676e+15 e *yval* igual a 35936060; o segundo (nó 3, AF>0.5) é uma folha composta por 17 observações com um valor previsto de 141168900 KG para a quantidade de cimento exportada e com *deviance* igual a 1.585472e+17. O nó 2 é ainda dividido no nó 4 (LOAm<139.81) e no nó 5 (LOAm>=139.81). O nó 4 é um nó folha composto por 9 observações com um valor previsto de 20502290 KG para a quantidade de cimento exportada e com *deviance* igual a 2.383056e+15. O nó 5 é um nó folha composto por 12 observações com um valor previsto de 47511380 KG para a quantidade de cimento exportada e com *deviance* igual a 1.728951e+15. Note-se que só 2 dos 6 preditores são de facto utilizados na construção da árvore de regressão: AF e LOAm.

A *package rpart.plot* permite-nos representar a árvore graficamente como se pode observar na figura 4.5.

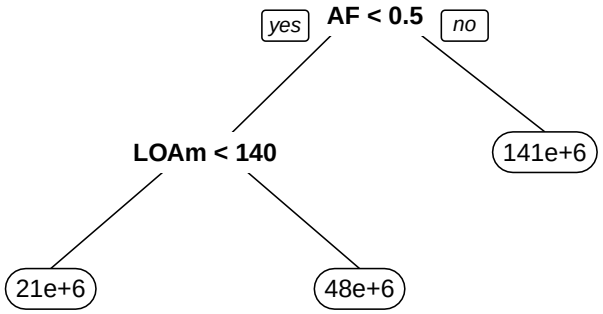

Figura 4.5: Árvore de regressão obtida através da *package rpart*.

Após a obtenção da árvore pode-se então analisar a evolução do erro da validação cruzada em função do parâmetro de complexidade (ou equivalentemente, em função da dimensão da árvore), e determinar o valor do parâmetro de complexidade para o qual a árvore parece ótima. A figura 4.6 mostra a variação do erro da validação cruzada em função do valor do parâmetro de complexidade.

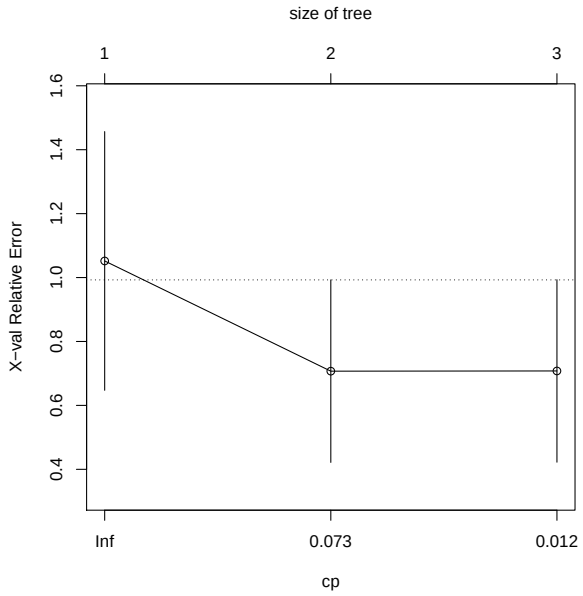

Figura 4.6: Gráfico do erro da Validação Cruzada contra o valor de cp.

Analogamente, usando a instrução *printcp()* pode-se obter essa informação, conduzindo ao seguinte *output*:

```
Regression tree:
rpart(formula = QUANT ~ ., data = DadosCimento, method = "anova")

Variables actually used in tree construction:
[1] AF   LOAm

Root node error: 2.7045e+17/38 = 7.117e+15

n= 38

        CP nsplit rel error  xerror    xstd
1 0.384684      0   1.00000 1.05196 0.40478
```

```
2 0.013872      1   0.61532 0.70710 0.28555
3 0.010000      2   0.60144 0.70778 0.28552
```

Os resultados anteriores exibem informação sobre a precisão do poder preditivo da árvore. Dá-nos acesso a dois tipos de erro: a coluna rel error indica-nos o erro relativo para as predições geradas que não têm qualquer utilidade para decidir o tamanho da árvore e a coluna xerror representa o erro da validação cruzada. Este último é uma medida útil para avaliar o desempenho da predição da variável em estudo. A coluna xstd representa o desvio padrão associado ao erro da validação cruzada.

De acordo com a metodologia seguida na construção da árvore, usando o *rpart* (ver expressão 3.10) que consiste em usar o parâmetro de complexidade que minimize o grau de complexidade, conduz a considerar 2 divisões, resultando em 3 nós folha, (compatível com a figura 4.6). Adicionalmente pode-se observar que o erro da validação cruzada é $70,8\%$ - relativamente elevado. Quando efetuamos a poda da árvore (usando o valor de cp=0.01) obtemos o mesmo resultado.

4.4.2 Aplicação da *package tree*

Utilizando a função *tree*, de uma forma similar, obtém-se o seguinte resultado:

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 38 2.704e+17  83010000
   2) COTA < -11.5 6 3.614e+16 248000000 *
   3) COTA > -11.5 32 4.033e+16  52080000
     6) AF < 0.5 21 7.864e+15  35940000
      12) GTm < 4536.25 5 4.940e+14   9997000 *
      13) GTm > 4536.25 16 2.954e+15  44040000 *
     7) AF > 0.5 11 1.655e+16  82890000 *
```

A raíz da árvore tem 38 observações e é dividida a partir da variável COTA. Obtém-se o nó 2 (COTA<-11.5) e o nó 3 (COTA>-11.5). O nó 2 é um nó folha com 6 observações com um valor previsto de 248000000 KG para a quantidade de cimento exportada e com *deviance* igual a 3.614e+16. O nó 3 com 32 observações é dividido pela variável AF no nó 6(AF<0.5) e no nó 7(AF>0.5). Por sua vez, o nó 6 é divido em duas folhas: a folha 12

(GTm<4536.25) e na folha 13 (GTm>4536.25). A folha 12 é composta por 5 observações com um valor previsto de 9997000 KG para a quantidade de cimento exportada e com *deviance* igual a 4.940e+14. A folha 13 é composta por 16 observações com um valor previsto de 44040000 KG para a quantidade de cimento exportada e com *deviance* igual a 2.954e+15. O nó 7 é uma folha composta por 11 observações com um valor previsto de 82890000 KG para a quantidade de cimento exportada e *deviance* igual a 1.655e+16. Neste caso são utilizados para a construção da árvore 3 dos 6 preditores: COTA, AF e GTm.

Através da função *plot.tree* obtém-se a figura 4.7 que representa a árvore graficamente.

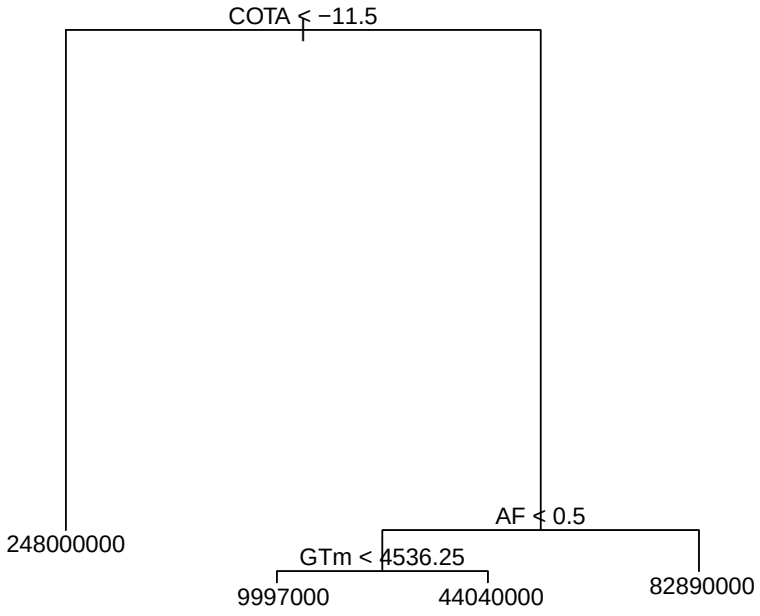

Figura 4.7: Árvore de regressão, sem poda, obtida através da *package tree*.

Árvore construída através da Poda

É do nosso interesse podar a árvore obtida. Para isso é importante verificar qual o tamanho da árvore que a otimiza. A partir do método da validação cruzada pode-se através da figura 4.8 escolher o melhor tamanho da árvore.

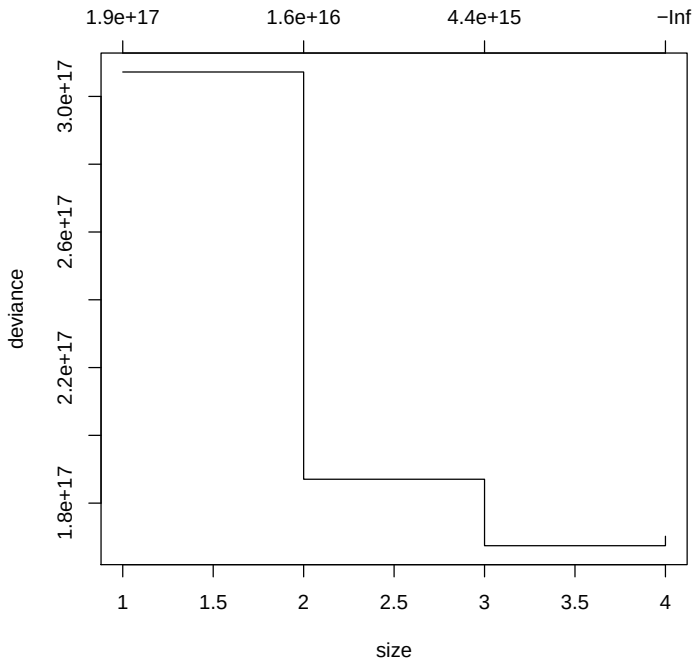

Figura 4.8: Gráfico da Validação Cruzada para a escolha da Complexidade da Árvore.

Note-se que o valor da *deviance* é menor quando o tamanho da árvore é 3. Por isso, efetuamos a poda da árvore com a pré-definição do *best*=3. Obtém-se o seguinte resultado:

```
node), split, n, deviance, yval
      * denotes terminal node

1) root 38 2.704e+17  83010000
  2) COTA < -11.5 6 3.614e+16 248000000 *
  3) COTA > -11.5 32 4.033e+16  52080000
    6) AF < 0.5 21 7.864e+15  35940000 *
    7) AF > 0.5 11 1.655e+16  82890000 *
```

A raíz da árvore tem 38 observações e é dividida a partir da variável COTA. Obtém-se o nó 2 (COTA<-11.5) e o nó 3 (COTA>-11.5). O nó 2 é um nó folha com 6 observações com um valor previsto de 248000000 KG para a quantidade de cimento exportada e com *deviance* igual a 3.614e+16. O nó 3 com 32 observações é dividido pela variável AF no nó 6(AF<0.5) e no nó 7(AF>0.5). O nó 6 é um nó folha com 21 observações com um valor previsto de 35940000 KG para a quantidade de cimento exportada e com *deviance* igual a

7.864e+15. O nó 7 é um nó folha com 11 observações com um valor previsto de 82890000 KG para a quantidade de cimento exportada e com *deviance* igual a 1.655e+16.

Através da função *plot.tree* obtém-se a figura 4.9 que representa a árvore podada graficamente.

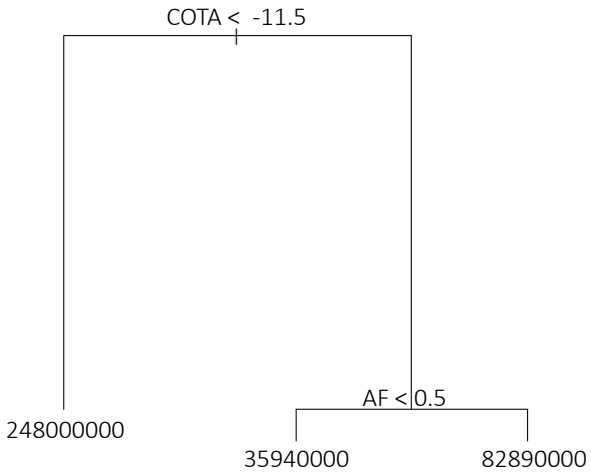

Figura 4.9: Árvore de regressão obtida através da *package tree* depois de efetuada a poda.

4.5 Análise dos Resultados Obtidos

Aquando da aplicação dos Modelos de Regressão Linear Múltipla verificou-se que as variáveis significativas para explicar a variação existente na variável resposta - raizQUANT - são: GTm, LOAm, CALm, C1, C2 e AF. A relação encontrada é dada por:

$$\begin{aligned}\text{raizQUANT} =& 6925,236 + 1,023\text{GTm} - 142,437\text{LOAm} + 2465,09\text{CALm} \\ & - 6547,529\text{C1} - 6109,977\text{C2} + 1958,979\text{AF}.\end{aligned}$$

É possível concluir que as variáveis mediana da arqueação bruta, mediana do calado e a existência da ferrovia afetam positivamente a raiz quadrada da quantidade de cimento exportada, enquanto que a variável mediana do comprimento afeta negativamente. As variáveis mudas de referência relativamente à cota de serviço máxima reforçam que a quantidade de cimento exportada aumenta com a profundidade máxima do canal principal de navegação.

No entanto, a predição de novos valores é complicada pois estar-se-ia a predizer valores para a raiz quadrada da quantidade de cimento exportada e não para a quantidade de cimento exportada. Na tentativa de combater este problema aplicou-se ao caso de estudo as Árvores de Regressão.

Quando fazemos a aplicação da *package rpart* pode-se sintetizar a informação obtida anteriormente na seguinte tabela:

		Previsão da Quantidade de Cimento	SSE	MSE*
	Existir Ferrovia	141168900 Kg	1,585472e+17	9,326303e+15
Não existir Ferrovia	LOAm < 139,81	20502290 Kg	2,383056e+15	2,64784e+14
	LOAm >= 139,81	47511380 kg	1,728951e+15	1,440793e+14

Tabela 4.12: Síntese da informação obtida, usando a *package rpart*.

* estes valores são obtidos através da instrução *summary*, que está no apêndice F e que dá informação mais detalhada.

Por outro lado, quando se aplica a *package tree* temos a seguinte tabela sumária:

		Previsão da Quantidade de Cimento	SSE	MSE
	COTA = -12,5	248000000 Kg	3,614e+16	6,023333e+15
COTA = -10,5 ou COTA = -8,5	Não existir Ferrovia	35940000 Kg	7,864e+15	3,744762e+14
	Existir Ferrovia	82890000 Kg	1,655e+16	1,504545e+14

Tabela 4.13: Síntese da informação obtida, usando a *package tree*.

Comparando as árvores obtidas pelas duas *packages* é fácil de observar a variável que têm em comum: AF. Isto significa que a existência da ferrovia tem uma importância acrescida na predição dos valores da quantidade de cimento exportada. Quando se utiliza a *package rpart* se não existir ferrovia a predição vai depender da mediana do comprimento do navio (LOAm). Quando se utiliza a *package tree* a divisão é efetuada a partir da variável COTA. Caso a COTA seja igual a -12,5 a árvore é divida através da variável AF. Se não existir ferrovia existe ainda mais uma divisão a partir da mediana da arqueação bruta (GTm). Quando se efetua a poda da árvore a última divisão é eliminada.

No entanto, quando se analisam os resultados em termos da quantidade de cimento exportado, verifica-se que a profundidade do canal é determinante para o aumento da exportação da quantidade de cimento. Basta reparar que neste caso a quantidade média prevista para a quantidade de cimento exportada, se a profundidade corresponder ao valor máximo, atinge o valor de 248000000 Kg (resultado obtido pela *package tree*). Esta metodologia de divisão corresponde ao maior valor médio obtido para a quantidade de cimento.

Por outro lado, pela *package rpart*, verifica-se que a maior quantidade de cimento exportada, em termos médios, é conseguida se existir a Ferrovia em funcionamento, correspondendo-lhe um valor 141168900 Kg - inferior ao anteriormente referido - mas correspondendo ao dobro do obtido no caso de não existir Ferrovia.

Em suma, sem dúvida que a profundidade da cota e o funcionamento da ferrovia são variáveis que contribuem de uma forma determinante para a quantidade de cimento exportada, quer usando a análise de regressão linear múltipla quer as árvores de regressão.

Capítulo 5

Conclusões

Neste último capítulo sumarizam-se alguns comentários finais acerca do estágio e sobre o estudo proposto pela organização de acolhimento - APA, S.A.

O estágio realizado no Porto de Aveiro permitiu adquirir conhecimento sobre a realidade portuária, potenciada pelas atividades diárias e pelo contacto com toda a sua envolvente. O estágio foi bastante enriquecedor a nível profissional e pessoal. Possibilitou-me um enquadramento num ambiente organizacional que resultou numa melhor compreensão do mundo empresarial. O Dr. Luís Sousa facultou todas as ferramentas necessárias para o desenvolvimento das tarefas diárias e também para o estudo efetuado.

O estudo proposto pelo Porto de Aveiro focou-se na exportação do cimento. O principal objetivo deste estudo foi tentar perceber que variáveis mais influenciam a quantidade de cimento exportada. Para alcançar este objetivo foram utilizadas duas metodologias: Modelos de Regressão Linear Múltipla e Árvores de Regressão. A aplicação destas metodologias ao caso de estudo proporcionou o desenvolvimento de competências distintas, algumas nunca antes exploradas, alargando o grau de conhecimento acerca das mesmas e cimentando as já conhecidas durante o percurso académico.

Aquando da aplicação dos Modelos de Regressão Linear Múltipla surgiram diversos problemas na primeira abordagem, pois não é a quantidade de cimento exportada que depende linearmente das restantes variáveis mas sim a raiz quadrada dessa quantidade. Notou-se que as variáveis: mediana da arqueação bruta, mediana do comprimento, mediana do calado, cota de serviço máxima e a existência da ferrovia explicam grande parte da variação existente na raiz quadrada da quantidade de cimento exportada. O poder preditivo fica comprometido devido ao facto de ser a raiz quadrada da quantidade de cimento exportada

a depender linearmente das variáveis descritas acima. De forma a combater essa falha aplicaram-se as Árvores de Regressão ao problema em questão.

A aplicação das Árvores de Regressão demonstra-nos que a existência da ferrovia influencia positivamente a predição de valores futuros sobre a quantidade de cimento exportada. As variáveis mediana do comprimento e cota de serviço máxima são também importantes para a predição.

Para a aplicação dos Modelos de Regressão Linear Múltipla foi utilizado o *software* SPSS e para a aplicação das Árvores de Regressão o *software* R.

Este caso de estudo demonstrou a suspeita já existente por parte do Porto de Aveiro, isto é, conclui-se que utilizando ambas as metodologias a existência da ferrovia influencia em muito a variação da quantidade de cimento exportada e também a predição de valores futuros. Notou-se também que a cota de serviço máxima é bastante significativa e que o Porto de Aveiro deve ter como objetivo a manutenção e dragagem do canal de navegação principal.

No entanto, existem algumas limitações em relação ao problema estudado: é necessário a recolha de dados mais pormenorizados por parte do Porto de Aveiro para que outras variáveis, consideradas importantes na envolvente portuária, possam ser consideradas no estudo; a qualidade dos dados deve ser averiguada diariamente para que a integridade dos estudos futuros não seja comprometida. O setor portuário não é muito explorado estatisticamente, mas com a recolha de informação mais detalhada e o apuramento da sua qualidade é possível uma exploração mais profícua a médio longo prazo.

Bibliografia

[1] Amorim, I. (2008). *PORTO DE AVEIRO: Entre a Terra e o Mar*. APA - Administração do Porto de Aveiro, S.A. Aveiro, Portugal.

[2] Arruda, C. M., Júnior, E. F. N., Magalhães, P. S. B. (2008). *Método dos Indicadores de Desempenho proposto pela ANTAQ: uma Aplicação ao Terminal Portuário de Pecém*. Rio de Janeiro, RJ, Brasil.

[3] Box, G.E.P. e Cox, D.R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol.26, No.2. pp. 211-252.

[4] Box, G.E.P. e Tidwell, P.W. (1962). Transformation of the independent variables. *Technometrics* 4, 531-550.

[5] Breiman, L., J.H. Friedman, R. A. Olshen, and C.G. Stone. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, USA.

[6] Clark, L. A., e D. Pregibon. (1992). *Tree-based Models*. J. M. Chambers and T.J. Hastie, editors, page 377-420.

[7] Crown, W. H. (1998). *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-dependent Variable Models*. Greenwood Publishing.

[8] Fonseca, J. (2001). *Estatística Matemática*, Vol. II Edições Sílabo.

[9] Glenn, A., Fabricius, K.E. (2000). Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology*, 81(11), pp. 3178-3192.

[10] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L. (2009). *Análise Multivariada de dados*, 6ª edição. Bookman.

[11] Hall, A., Neves, C. e Pereira, A. (2011). *Grande Maratona de Estatística no SPSS*. Escolar Editora.

[12] Koenker, R. (1981). A Note on Studentizing a Test for Heterocedasticity. *Journal of Econhometrics*, 17, 107-112.

[13] Kranowsky, A.W. (1998). *An Introduction to Statistical Modelling*. John Wiley & Sons, Ltd.

[14] Maindonald, J., Braun, W. J. (2010). *Data Analysis and Graphics Using R: An Example-Based Approach Third Edition.* Cambridge University Press.

[15] Miranda, M.M.S. (2012). Apontamentos da cadeira de Modelos Estatísticos.

[16] Montgomery, D.C., Peck, E.A., Vining, G.G. (2001). *Introduction to Linear Regression Analysis Third Edition*. John Wiley & Sons, Inc.

[17] Murteira, B., Ribeiro, C.S., Andrade e Silva, J., Pimenta, C. (2010). *Introdução à Estatística*. Escolar Editora.

[18] Ott, R., Longnecker, M. (2008). *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning.

[19] Plano Estratégico do Porto de Aveiro 2006. APA - Administração do Porto de Aveiro, S.A., 2014.

[20] Portal da APA

[21] Portal da APFF

[22] Relatório de Sustentabilidade 2012. APA - Administração do Porto de Aveiro, S.A., 2014.

[23] Relatório de Sustentabilidade 2012. APFF - Administração do Porto da Figueira da Foz, S.A., 2014.

[24] Ripley, B.D. (1996). *Pattern Recognition and Neural Network*. Cambridge University Press, Cambrigde, UK.

[25] Ryan, T.P. (1997). *Modern regression methods*. Wiley.

[26] Sassi, C. P., Perez, F. G., Myazato, L., Ye, X., Ferreira-Silva, P. H., Louzada, F. *Modelos de Regressão linear Múltipla utilizando os softwares R e Statistica: Uma aplicação a dados de conservação de frutas*. ICMC – USP – CP668 – CEP 13.566-590, São Carlos, SP, Brasil.

[27] Sen, A.K., Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications*. Springer-Verlag.

[28] Wilcox, R. (2011). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. CRC Press.

Apêndice A

Valores Críticos do teste de Durbin-Watson

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-----	-----	-----	-----	-----	-----	-----	-----
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	-----	-----	-----	-----	-----	-----
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	-----	-----	-----	-----
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	-----	-----
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956

Tabela A.1: Valores Críticos do teste de Durbin-Watson.

Apêndice B

Dados

	QUANT	raizQUANT	GTm	LOAm	CALm	COTA	C1	C2	AF	PROD
1T/2005	952875	976,15	1749,5	77,5	4,13	-8	1	0	0	9898,95
2T/2005	12304476	3507,77	2472	91	5,1	-8	1	0	0	40539,53
3T/2005	9225692	3037,38	2999	92,5	5,99	-8	1	0	0	37398,57
4T/2005	16964857	4118,84	6167	131	8,03	-8	1	0	0	69218,25
1T/2006	27499648	5244,01	2575	86,04	5,55	-8	1	0	0	60043,19
2T/2006	43394885	6587,48	6124,5	122,18	7,09	-8	1	0	0	63685,52
3T/2006	40732284	6382,18	8254	142,91	7,22	-8	1	0	0	55668,36
4T/2006	0	0	0	0	0	-8	1	0	0	0
1T/2007	32851591	5731,63	8254	142,91	7,22	-8	1	0	0	61078,69
2T/2007	36850707	6070,48	8254	142,91	7,22	-8	1	0	0	51223,62
3T/2007	38392056	6196,13	8254	142,91	7,22	-8	1	0	0	68475,38
4T/2007	48277823	6948,22	6073,5	122,87	6,94	-8	1	0	0	70207,98
1T/2008	41936991	6475,88	8254	142,91	7,22	-8	1	0	0	65784,58
2T/2008	53454445	7311,25	7447,5	139,86	7,22	-8	1	0	0	62301,89
3T/2008	70655758	8405,7	8254	142,91	7,22	-8	1	0	0	65148,06
4T/2008	50659647	7117,56	8254	142,91	7,22	-8	1	0	0	75981,8
1T/2009	25900382	5089,24	8254	142,91	7,22	-10,5	0	1	0	74991,43
2T/2009	37330532	6109,87	8254	142,91	7,22	-10,5	0	1	0	63687,06
3T/2009	53453174	7311,17	8254	142,91	7,22	-10,5	0	1	0	195499,33
4T/2009	70018759	8367,72	8254	142,91	7,22	-10,5	0	1	0	115380,57
1T/2010	43800664	6618,21	8254	142,91	7,22	-10,5	0	1	0	78304,19
2T/2010	88034203	9382,65	4345	112,95	6,9	-10,5	0	1	1	86051,91
3T/2010	59132595	7689,77	8254	142,91	7,22	-10,5	0	1	1	93374,99
4T/2010	54435480	7378,04	1945	87,97	6,59	-10,5	0	1	1	75357,92
1T/2011	62148751	7883,45	2545	88,6	6,42	-10,5	0	1	1	60901,84
2T/2011	57966798	7613,59	3881	112,78	6,7	-10,5	0	1	1	72795,14
3T/2011	59538374	7716,11	8254	142,91	7,22	-10,5	0	1	1	66263,38
4T/2011	34476042	5871,63	8254	142,91	7,22	-10,5	0	1	1	59980,69
1T/2012	74703369	8643,11	8254	142,91	7,22	-10,5	0	1	1	97510,54
2T/2012	163425330	12783,79	5581	108,2	7,06	-10,5	0	1	1	84240,85
3T/2012	114629463	10706,51	7078	125,86	7,22	-10,5	0	1	1	86927,92
4T/2012	143284551	11970,15	6204	117	7,22	-10,5	0	1	1	56247,06
1T/2013	186713522	13664,32	5125,5	110,05	6,89	-12,5	0	0	1	76087,86
2T/2013	238830825	15454,15	5629	119,95	7,06	-12,5	0	0	1	68995,56
3T/2013	187805644	13704,22	7617	125	7,22	-12,5	0	0	1	99748,68
4T/2013	182924982	13524,98	5164	113,47	6,88	-12,5	0	0	1	91408,81
1T/2014	294597024	17163,83	5687,5	116,58	7,14	-12,5	0	0	1	75836,36
2T/2014	397225127	19930,51	5629	114,44	7,05	-12,5	0	0	1	93682,49

Tabela B.1: Dados utilizados.

Apêndice C

Resultados

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,903[a]	,815	,772	4,07846E7	1,166

a. Predictors: (Constant), c2, Mediana da Arqueação Bruta por trimestre, Ferrovia, Produtividade média de atendimento, Mediana do Calado por trimestre, c1, Mediana do Comprimento por trimestre

b. Dependent Variable: Quantidade de mercadoria movimentada por trimestre

Tabela C.1: Resumo do modelo.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1,614E8	5,073E7		3,182	,003	5,781E7	2,650E8		
	Mediana da Arqueação Bruta por trimestre	17722,894	10301,561	,495	1,720	,096	-3315,701	38761,489	,074	13,471
	Mediana do Comprimento por trimestre	-2607757,751	1405585,703	-,873	-1,855	,073	-5478346,718	262831,215	,028	36,038
	Mediana do Calado por trimestre	3,418E7	1,808E7	,526	1,891	,068	-2737579,424	7,110E7	,080	12,573
	Ferrovia	3,290E7	2,624E7	,194	1,254	,220	-2,068E7	8,648E7	,257	3,888
	Produtividade média de atendimento	167,771	327,816	,059	,512	,613	-501,719	837,261	,460	2,172
	c1	-1,523E8	3,444E7	-,891	-4,421	,000	-2,226E8	-8,194E7	,151	6,606
	c2	-1,534E8	2,151E7	-,898	-7,132	,000	-1,974E8	-1,095E8	,388	2,577

Tabela C.2: Estimação dos coeficientes, intervalos de confiança e teste da colinearidade.

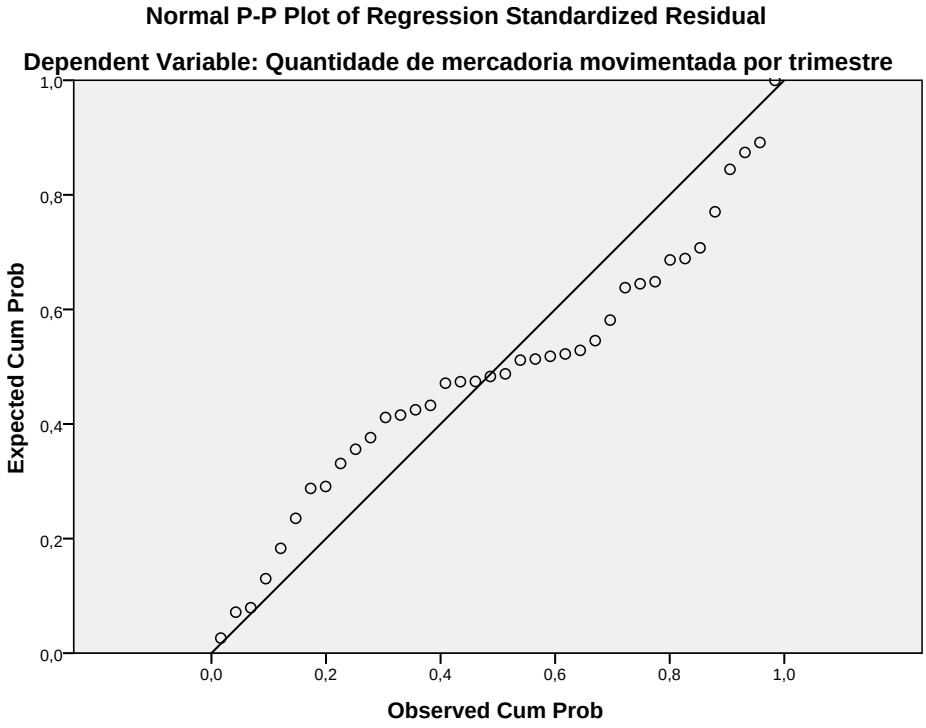

Figura C.1: *PP-plot* (Normal) dos resíduos.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,141	38	,056	,881	38	,001

a. Lilliefors Significance Correction

Tabela C.3: Resultado dos testes de Kolmogorov-Smirnov e de Shapiro-Wilk.

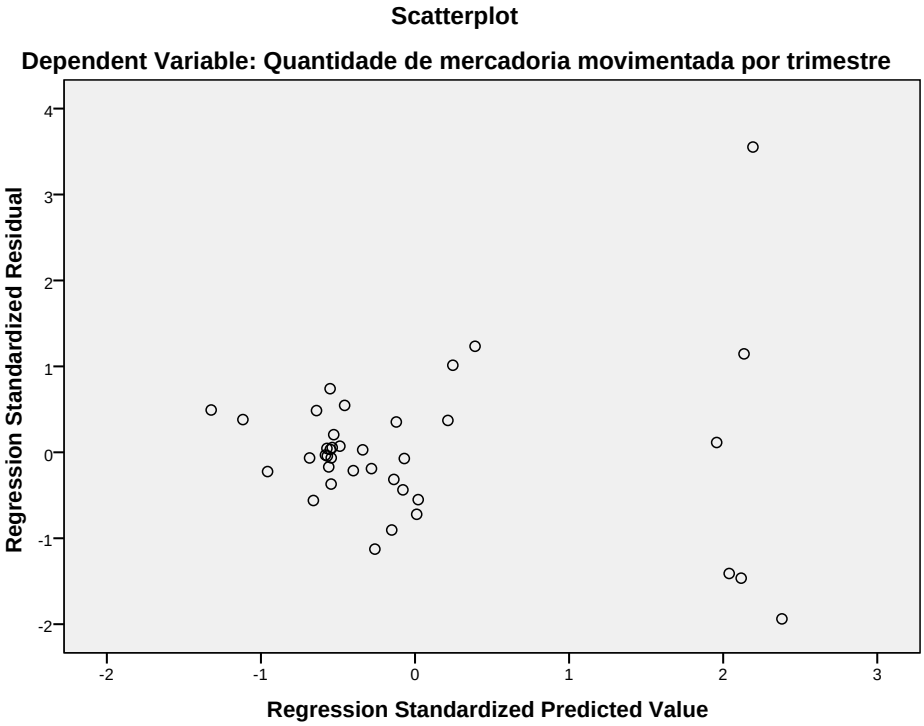

Figura C.2: Gráfico de resíduos versus valores preditos.

Apêndice D

MACRO

```
* BREUSCH-PAGAN & KOENKER TEST MACRO *
* See 'Heteroscedasticity: Testing and correcting in SPSS'
* by Gwilym Pryce, for technical details.
* Code by Marta Garcia-Granero 2002/10/28.

* The MACRO needs 3 arguments:
* the dependent, the number of predictors and the list of predictors
* (if they are consecutive, the keyword TO can be used) .

* (1) MACRO definition (select an run just ONCE).

DEFINE bpktest(!POSITIONAL !TOKENS(1) /!POSITIONAL !TOKENS(1)
 /!POSITIONAL !CMDEND).
* Regression to get the residuals and residual plots.
REGRESSION
/STATISTICS R ANOVA
/DEPENDENT !1
/METHOD=ENTER !3
/SCATTERPLOT=(*ZRESID,*ZPRED)
/RESIDUALS HIST(ZRESID) NORM(ZRESID)
/SAVE RESID(residual) .
do if $casenum=1.
print /"Examine the scatter plot of the residuals to detect"
```

```
/"model misspecification and/or heteroscedasticity"
/""
/"Also, check the histogram and np plot of residuals "
/"to detect non normality of residuals "
/"Skewness and kurtosis more than twice their SE indicate non-normality ".
end if.
* Checking normality of residuals.
DESCRIPTIVES
VARIABLES=residual
/STATISTICS=KURTOSIS SKEWNESS .
* New dependent variable (g) creation.
COMPUTE sq_res=residual**2.
compute constant=1.
AGGREGATE
/OUTFILE='tempdata.sav'
/BREAK=constant
/rss = SUM(sq_res)
/N=N.
MATCH FILES /FILE=*
/FILE='tempdata.sav'.
EXECUTE.
if missing(rss) rss=lag(rss,1).
if missing(n) n=lag(n,1).
compute g=sq_res/(rss/n).
execute.
* BP&K tests.
* Regression of g on the predictors.
REGRESSION
/STATISTICS R ANOVA
/DEPENDENT g
/METHOD=ENTER !3
/SAVE RESID(resid) .
*Final report.
do if $casenum=1.
```

```
print /" BP&K TESTS"
/" ==========".
end if.
* Routine adapted from Gwilym Pryce.
matrix.
compute p=!2.
get g /variables=g.
get resid /variables=resid.
compute sq_res2=resid&**2.
compute n=nrow(g).
compute rss=msum(sq_res2).
compute ii_1=make(n,n,1).
compute i=ident(n).
compute m0=i-((1/n)*ii_1).
compute tss=transpos(g)*m0*g.
compute regss=tss-rss.
print regss
/format="f8.4"
/title="Regression SS".
print rss
/format="f8.4"
/title="Residual SS".
print tss
/format="f8.4"
/title="Total SS".
compute r_sq=1-(rss/tss).
print r_sq
/format="f8.4"
/title="R-squared".
print n
/format="f4.0"
/title="Sample size (N)".
print p
/format="f4.0"
```

```
/title="Number of predictors (P)".
compute bp_test=0.5*regss.
print bp_test
/format="f8.3"
/title="Breusch-Pagan test for Heteroscedasticity"
+ " (CHI-SQUARE df=P)".
compute sig=1-chicdf(bp_test,p).
print sig
/format="f8.4"
/title="Significance level of Chi-square df=P (H0:"
+ "homoscedasticity)".
compute k_test=n*r_sq.
print k_test
/format="f8.3"
/title="Koenker test for Heteroscedasticity"
+ " (CHI-SQUARE df=P)".
compute sig=1-chicdf(k_test,p).
print sig
/format="f8.4"
/title="Significance level of Chi-square df=P (H0:"
+ "homoscedasticity)".
end matrix.
!ENDDEFINE.
```

Apêndice E

Resultados *Stepwise*

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,680[a]	,462	,447	3109,72718	
2	,752[b]	,566	,541	2833,05838	
3	,828[c]	,685	,657	2448,39353	
4	,913[d]	,833	,813	1809,30871	1,191

a. Predictors: (Constant), Ferrovia

b. Predictors: (Constant), Ferrovia, Mediana do Calado por trimestre

c. Predictors: (Constant), Ferrovia, Mediana do Calado por trimestre, c2

d. Predictors: (Constant), Ferrovia, Mediana do Calado por trimestre, c2, c1

e. Dependent Variable: raizQUANT

Tabela E.1: Resumo do modelo.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	5600,328	678,598		8,253	,000	4224,067	6976,589		
	Ferrovia	5639,720	1014,565	,680	5,559	,000	3582,086	7697,354	1,000	1,000
2	(Constant)	-1198,694	2429,410		-,493	,625	-6130,659	3733,271		
	Ferrovia	5112,764	942,066	,616	5,427	,000	3200,268	7025,260	,963	1,039
	Mediana do Calado por trimestre	1044,549	360,949	,329	2,894	,007	311,785	1777,314	,963	1,039
3	(Constant)	-1585,595	2102,322		-,754	,456	-5858,027	2686,836		
	Ferrovia	6334,459	882,550	,763	7,177	,000	4540,901	8128,018	,819	1,221
	Mediana do Calado por trimestre	1221,235	315,807	,384	3,867	,000	579,439	1863,032	,939	1,065
	c2	-3205,249	893,748	-,384	-3,586	,001	-5021,564	-1388,934	,810	1,234
4	(Constant)	6151,042	2111,671		2,913	,006	1854,816	10447,268		
	Ferrovia	2397,458	977,272	,289	2,453	,020	409,183	4385,733	,365	2,741
	Mediana do Calado por trimestre	997,893	236,998	,314	4,211	,000	515,716	1480,070	,911	1,098
	c2	-6656,588	918,310	-,797	-7,249	,000	-8524,904	-4788,272	,419	2,386
	c1	-7167,753	1325,069	-,858	-5,409	,000	-9863,626	-4471,880	,201	4,968

a. Dependent Variable: raizQUANT

Tabela E.2: Estimação dos coeficientes, intervalos de confiança e teste da colinearidade.

Apêndice F

Código de Implementação

```
### Aplicação dos Dados às Árvores de Regressão

## Leitura dos dados

dados<-read.table("DadosCimentoR.txt", header=T, dec=",")
dados
attach(dados)

DadosCimento<-data.frame(subset(dados,select=c(GTm,LOAm,CALm,
COTA,AF,PROD,QUANT)))
DadosCimento
attach(DadosCimento)

## Aplicação utilizando a package rpart

library(rpart)

# Criação da árvore utilizando o método anova
arv.cimento <- rpart(QUANT~., method="anova", data=DadosCimento)

# Apresenta os resultados
print(arv.cimento)
```

```
# Fornece resultados detalhados
summary(arv.cimento)

library(rpart.plot)
rpart.plot(arv.cimento)

# A função printcp fornece informação para
# selecionarmos o valor adequado para cp.
printcp(arv.cimento)
# Faz o plot dos resultados da validação cruzada
plotcp(arv.cimento)

# Com o valor assim determinado para cp, obtém-se uma nova árvore podada.
prune(arv.cimento, cp=0.013872)

## Aplicação utilizando a package tree

library(tree)

# Criação da árvore
arvore.tree <- tree(QUANT~.,DadosCimento)

# Apresenta a árvore
arvore.tree

plot(arvore.tree); text(arvore.tree)

# Validação Cruzada
cv.tree(arvore.tree)
plot(cv.tree(arvore.tree)) # aqui podemos verificar qual o melhor best

prune.tree <- prune.tree(arvore.tree,best=3)
prune.tree
plot(prune.tree)
```